XTEL

Starting from…

(talking about the workplace)

What will be the biggest challenge to face,
what makes me most **uncertain** or **insecure**?

XTEL

Data platforms

The era of data-driven software, how it happened

How data has influenced software architecture over the last 40 years

2024 Q2

Glauco Uri

Who am I

glauco.uri@xtel-group.com
www.linkedin.com/in/glaucouri/

XTEL

XTEL is a leading Global Revenue Management Solutions Provider

XTEL delivers holistic, proven digitalization and change management for consumer packaged goods companies (CPGs) across all five continents. Combining state-of-the-art technology, deep industry expertise, data management and consulting capabilities, we can support CPG's of any size in their journey for value realization and profitable revenue growth.

https://www.xtel-group.com/

AGENDA

1 An overview

2 History of data platforms

3 The `big data` cultural revolution

4 A Modern ML data platform

5 Q&A

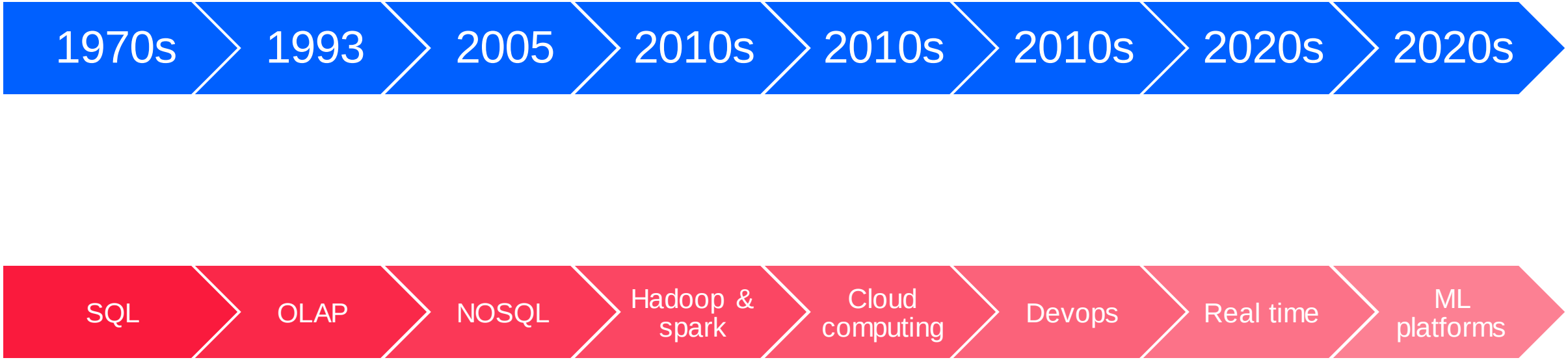
What is a Data Platform

A data platform is an integrated set of technologies that collectively meet an organization's end-to-end **data needs**. It enables the acquisition, storage, preparation, delivery, and governance of your data, as well as a security layer for users and applications.

A central processing hub for an organization's **data ecosystem**.

An overview

the most significant steps of modern platforms



SQL

1970s

SQL was invented in the 1970s based on the relational data model.

It was initially known as **S**tructured **E**nglish **QUE**ry **L**anguage [SEQUEL]

later shortened to SQL

```sql
SELECT
    first_name,
    last_name,
    roll_number,
    score,
    ((score / 50) * 100) as percentage
FROM students s
WHERE score > 50
ORDER BY
    percentage DESC, first_name, roll_number
```

SQL

1970s

Pros

- Ease of Learning and Use
- Data Integrity
- Structured Querying
- Scalability
- Multi-user Support
- Data Security
- Integration with Programming Languages
- Normalization for Data Organization
- Support for Constraints
- Extensive Community and Resources

Cons

- Limited Scalability for Complex Queries
- Can handle only known data
- Limit the shape of the data
- Non-compatibility with other Databases
- Steep Learning Curve for Advanced Features
- Vendor-specific Implementations
- Performance Issues
- Cost of Implementation and Maintenance
- Inflexibility
- Lack of scalability

The Rise of the OLAP Cube – aka data cube

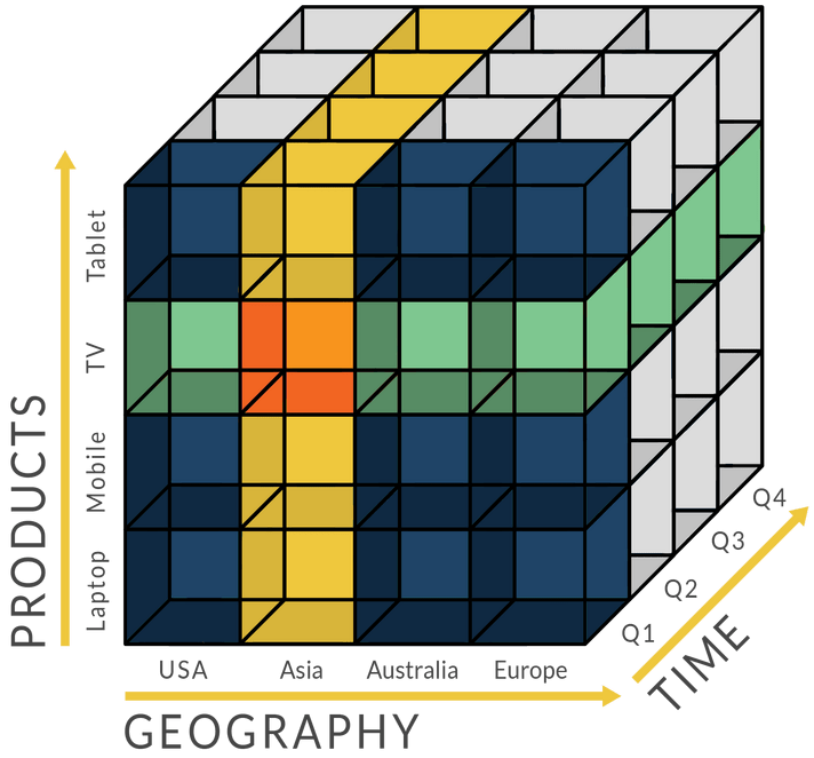
1990s

The term was invented by database legend Edgar F. Codd, in a 1993 paper titled Providing OLAP(On-Line Analytical Processing) to User-Analysts: An IT Mandate.

The OLAP cube grew out of a simple idea in computer programming.

OLAP databases introduce the **data denormalization** technique: pre-processed data. Storing pre-calculated aggregates eliminates the need for complex joins and speeds up analytical queries.

Often the da cube is created off-line, in a **data warehouse** storage.



The revolution of NOSQL databases

2000s

The acronym NoSQL was first used in late 90s by Carlo Strozzi while naming his lightweight, open-source “relational” database that did not use SQL.

The name came up again in 2009 used it to describe non-relational databases.

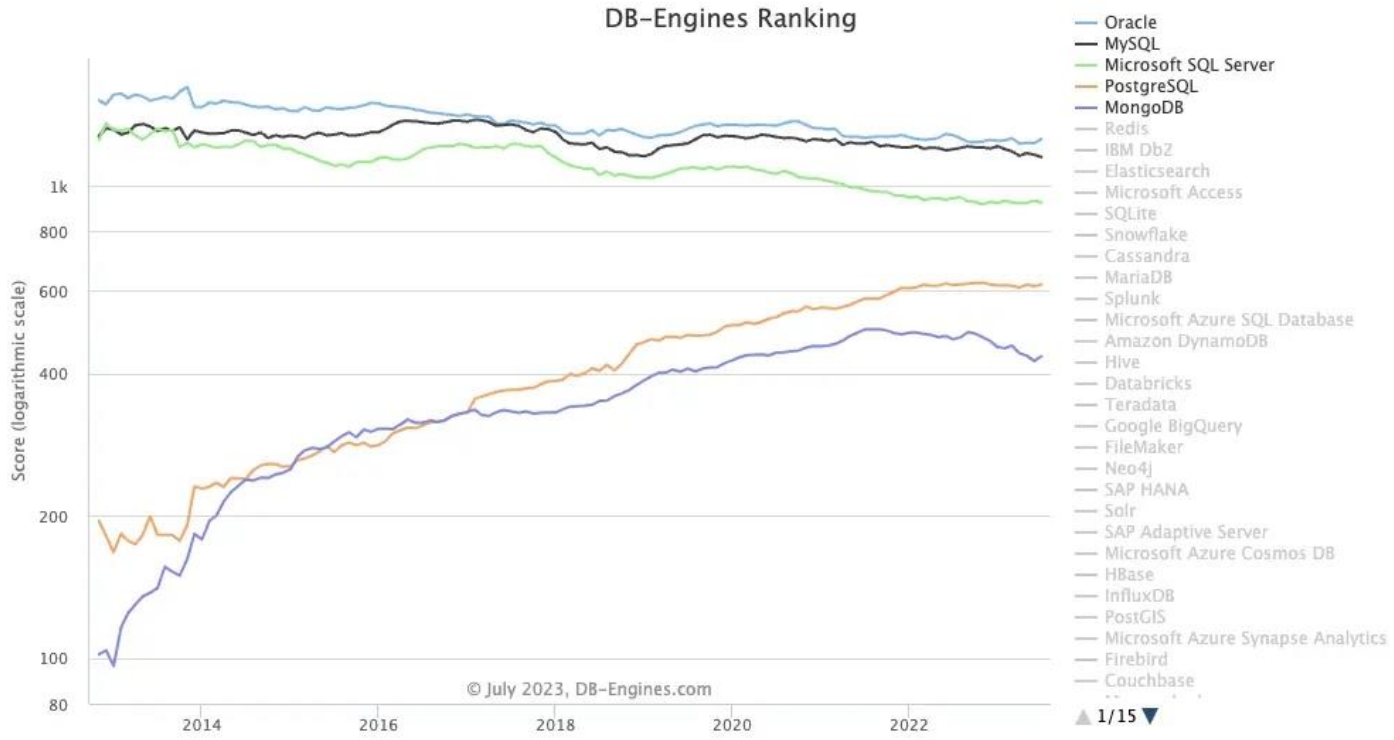
Relational databases are often referred to as SQL systems. The term NoSQL can mean either “No SQL systems” or the more commonly accepted translation of “Not only SQL,” to emphasize the fact some systems might support SQL-like query languages.

```
> db.user.find().pretty()
{
        "_id" : ObjectId("59b19f084597373d365674c0"),
        "firstname" : "WEI",
        "lastname" : "Tang",
        "email" : "weitang0326@gmail.com",
        "username" : "weitang0326@gmail.com",
        "password" : "mocha",
        "__v" : 0
}
```

The revolution of NOSQL databases

2000s

- Key-Value storage
- Document oriented [*]
- Column-oriented
- Graph database



[*] MongoDB:

- 10gen began developing MongoDB in 2007.
- In 2009, the company shifted to an open-source.
- In 2013, 10gen changed its name to **MongoDB**

Source: https://www.bytebase.com

Business description

CPG companies

Manufacturers | Distributors | Retailers | Consumers

Business description

CPG companies

Functional Goal	Levers	Decisions	Business Issues / Insights	KPIs	Data Sources
Increase Profitability	Product/ Packaging	Can we cost reduce the product formulation and/or packaging without alienating consumers?	Are there alienation or product issues with the reduced-cost product?	parity product performance vs. current-pi, attributes, satisfaction	product test, cost analysis
	Pricing	Can we increase product price without losing too much volume?	What is the price/volume trade-off at different prices?	price elasticity	price elasticity & optimization
	SKU Rationalization	Will the cost savings outweigh any loss through SKU rationalization?	What SKU's can we eliminate without alienating consumers?	substitutability, gain/loss analysis	SKU optimization
	Promotion	Can we decrease the number, frequency or value of promotions without losing consumers?	Can we maintain sales with fewer, less frequent, lower value promotions?	ROI/ contribution to sales, scenario simulations	marketing mix
	Cost Management	How should we respond to commodity cost increases?	Can we improve purchasing efficiency or pass costs on to consumers/channel?	direct ingredient cost, price elasticity, portfolio mix	R&D cost projections, price elasticity

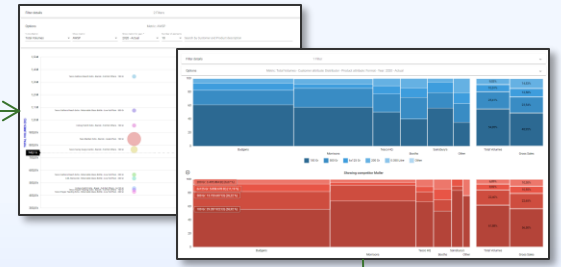
Business description **integrated solution** | RGM advanced analytics

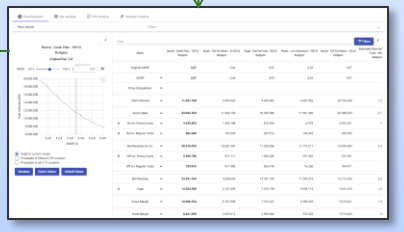
Visualize performance vs. targets and LY, by KPI, channel, customer & brand

Deep-dive into pricing and value chains at total level then drill down MD hierarchies

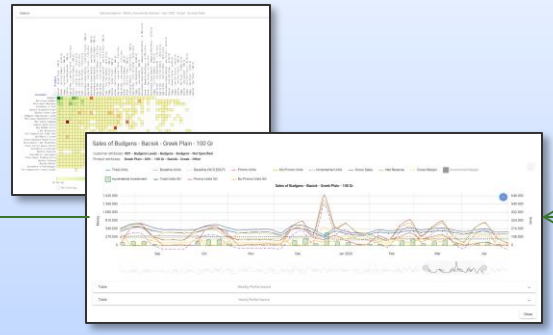
Establish defensibility across geographies, customers & brands

Investigate mix, price gaps, RTM share, etc. identifying risks and opportunities

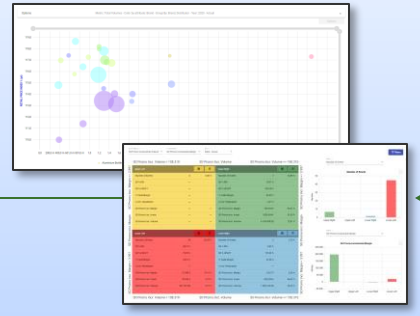




Model and simulate the effects of taking price, by pack and retailer, and a 'total' level, with AI generated elasticity curves for base and promo pricing



Deep dive into promo performance across KPI's and metrics, establishing the drivers of performance and view trends to identify pinch-points



Identify the best / worst performing promos by group and causal factors: e.g. duration, feature, display, etc.

Use AI/ML enabled constraint based modelling to optimize trade spend and investment levels by type to create optimal plans and promo guidelines

Promo guidelines and funds feed into the transactional modules to form the framework for KAMs activity plans

COFFEE BREAK

2000s The era of Big Data begins

- Industry digitalization
- Mobile
- IoT
- Web 2.0
- N-tier architecture

- in 1986, 99.2% of all storage capacity was analog
- in 2007, 94% of storage capacity was digital, a complete reversal of roles
- in 2002, digital information storage surpassed non-digital for the first time
- Multicultural environment

Source:
https://www.forbes.com/
http://www.martinhilbert.net/WorldInfoCapacity.html

Distributed computation with Hadoop + spark

2000s The era of Big Data

- There are mainly **two problems** with the big data.
  - First one is to store such a huge amount of data and the
  - Second one is to process that stored data.

Source: https://www.bytebase.com

Distributed computation with Hadoop + spark

2010s The era of Big Data

- The Hadoop was started by Doug Cutting and Mike Cafarella in 2002. Its origin was the Google File System paper, published by Google.
- In 2004, Google published one more paper on the technique **map-reduce**, which was the solution of processing those large datasets.

- In December of 2011, Apache Software Foundation released Apache Hadoop version 1.0.

- Spark was initially started at UC Berkeley's AMPLab in 2009, and open sourced in 2010.
- In 2013, the project was donated to the Apache Software Foundation.
- In February 2014, Spark became a Top-Level Apache Project.

- There are mainly **two problems** with the big data.
  - First one is to store such a huge amount of data and the
  - Second one is to process that stored data.

The traditional approach like RDBMS is not sufficient due to the heterogeneity of the data.

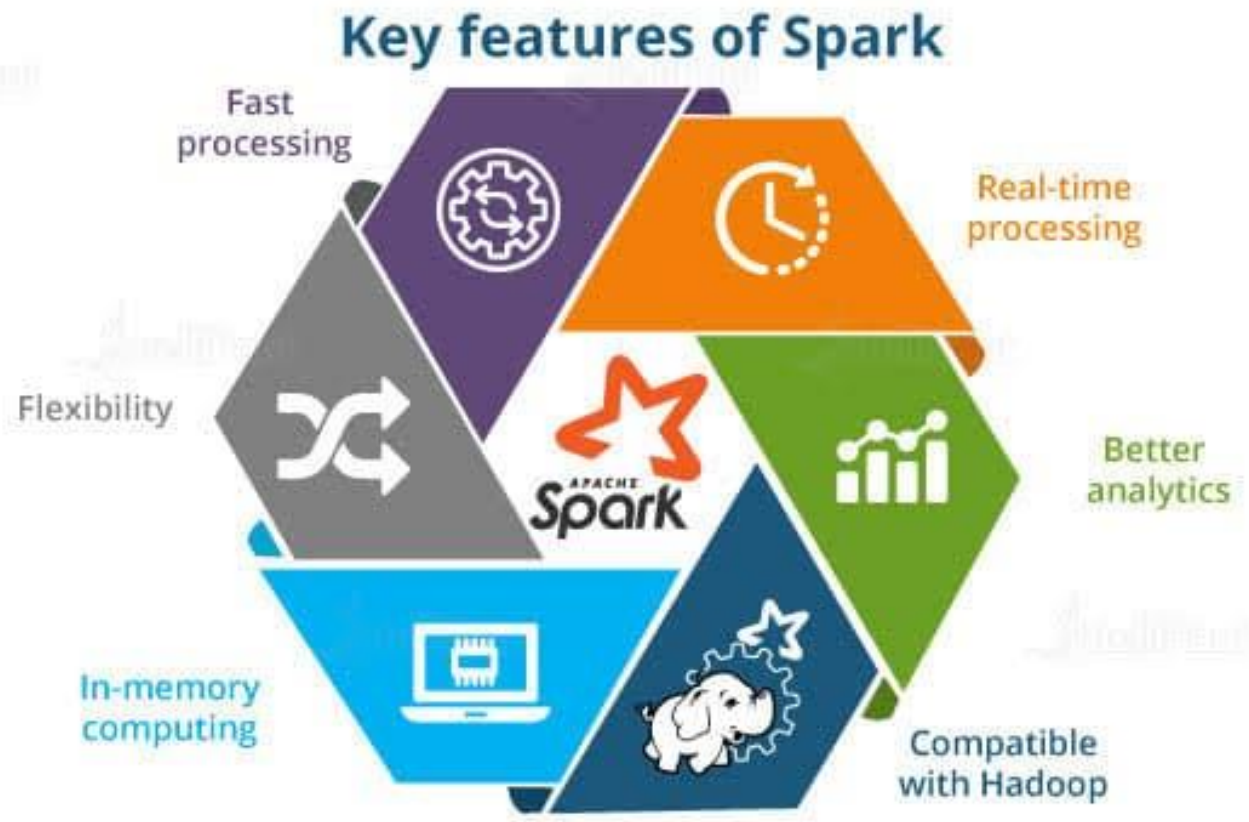
Hadoop comes as the solution to the problem of big data i.e. storing and processing the big data with some extra capabilities.

Spark is a unified analytics engine for large-scale data processing

Source: https://www.datascienceassn.org/

Distributed computation with Hadoop + spark

2010s The era of Big Data

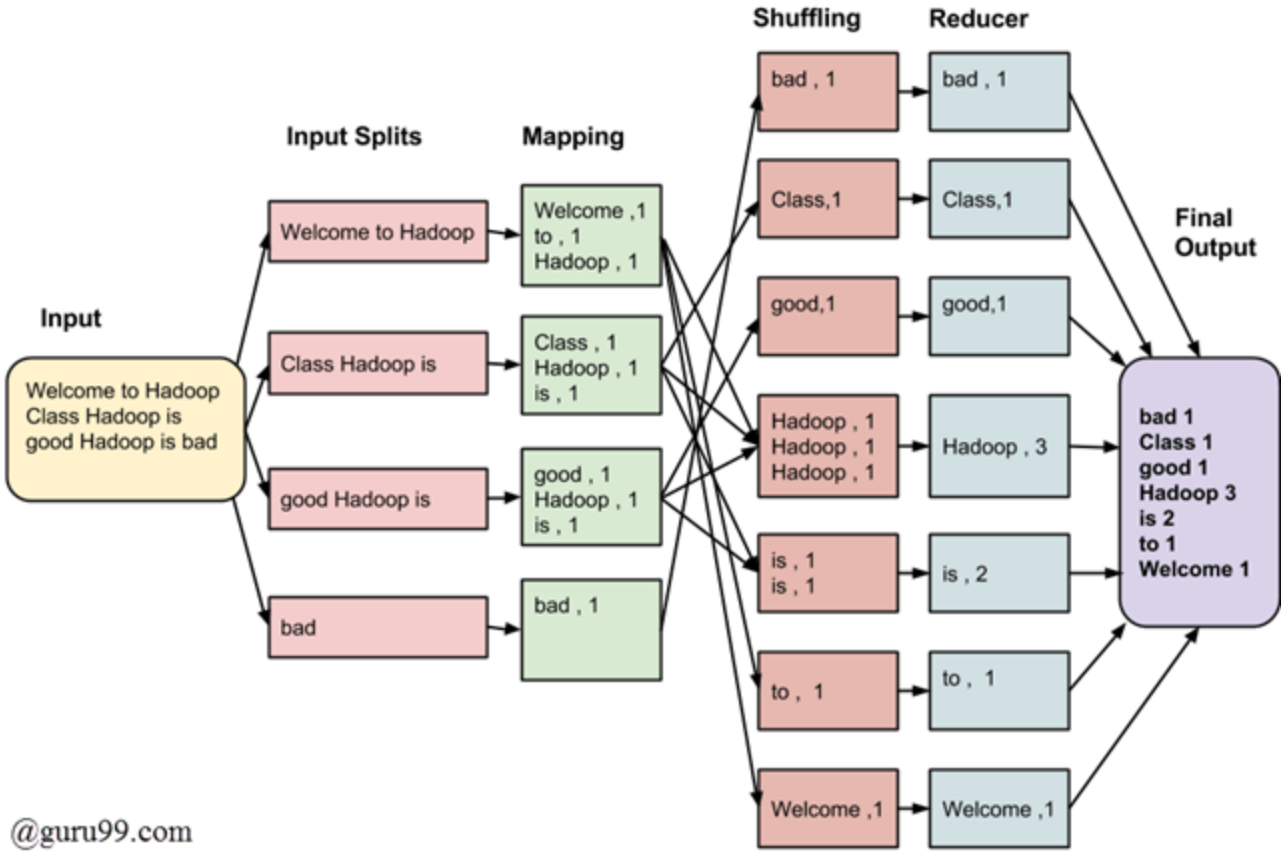


Source: https://intellipaat.com/

Distributed computation with Hadoop + spark

2010s The era of Big Data

Data locality in MapReduce refers to the ability to move the computation close to where the actual data resides on the node, instead of moving large data to computation.



Cloud computing

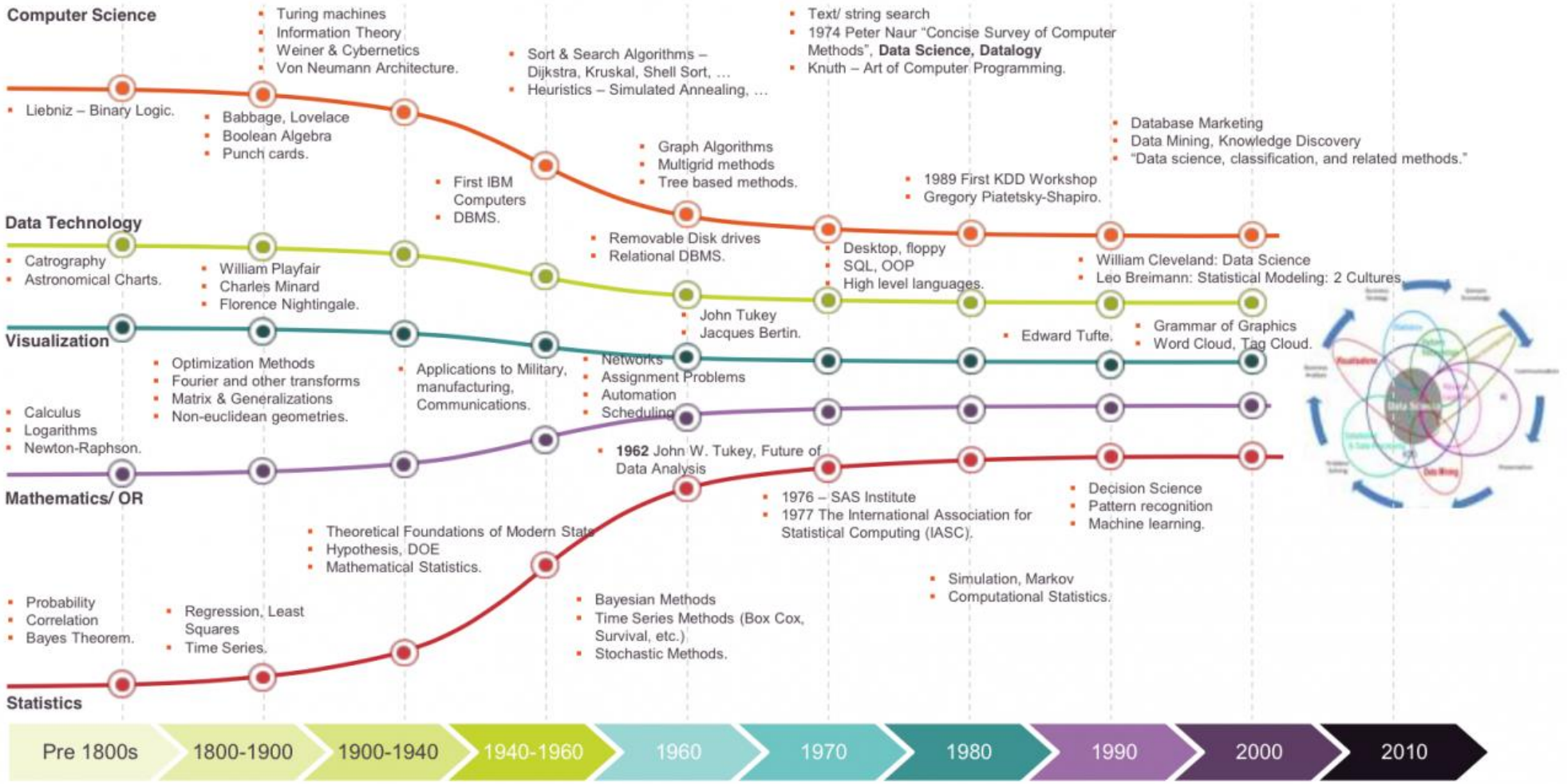
2010s

- The service-oriented architecture (SOA) promotes the idea of "Everything as a Service"
- Infrastructure as a service (IaaS)
- Platform as a service (PaaS)
- Software as a service (SaaS)
- Serverless computing or Function-as-a-Service (FaaS)

- New deployment models, software lifecycle, architecture
- New disciplines: Cloud engineering

2000s The era of Big Data begins

Source: https://www.datascienceassn.org/



Devops

2010s

DevOps is a methodology in the software development and IT industry. Used as a set of practices and tools, DevOps integrates and **automates** the work of software development (Dev) and IT operations (Ops) as a means for improving and shortening the systems development life cycle

- The History of DevOps can be traced back to the early 2000s when the Agile Software Development movement emerged
- The DevOps movement started to coalesce some time between 2007 and 2008, when IT operations and software development communities raised concerns what they felt was a fatal level of dysfunction in the industry.
- The first DevopsDays event!

Faster Time-to-Market: DevOps enables teams to deliver software quickly and efficiently.

Improved Collaboration: DevOps encourages collaboration between developers, testers, and IT operations teams, which leads to better communication and a faster resolution of issues.

Increased Quality: By **automating** testing and deployment processes, DevOps reduces the chance of errors and increases the overall software quality.CD/CI **pipelines**

Greater Efficiency: DevOps **automates repetitive tasks** and streamlines processes, freeing time and resources for more important tasks. Data **pipelines**

Source: https://www.atlassian.com/
https://en.wikipedia.org/

Real time computing

2020s

Real-time data is the process of analyzing data to **create insights in real time**. When raw data is received, it is immediately processed to empower near-instant decision-making. Instead of being stored, it is made available to promote insights as quickly as possible, furthering organizations' profitability, efficiency, and business outcomes

The 2020s have witnessed a significant shift in data engineering towards real-time processing and the integration of artificial intelligence (AI) and machine learning (ML) into data processing pipelines

Source: https://www.hpe.com/

Machine learning platforms

2020s

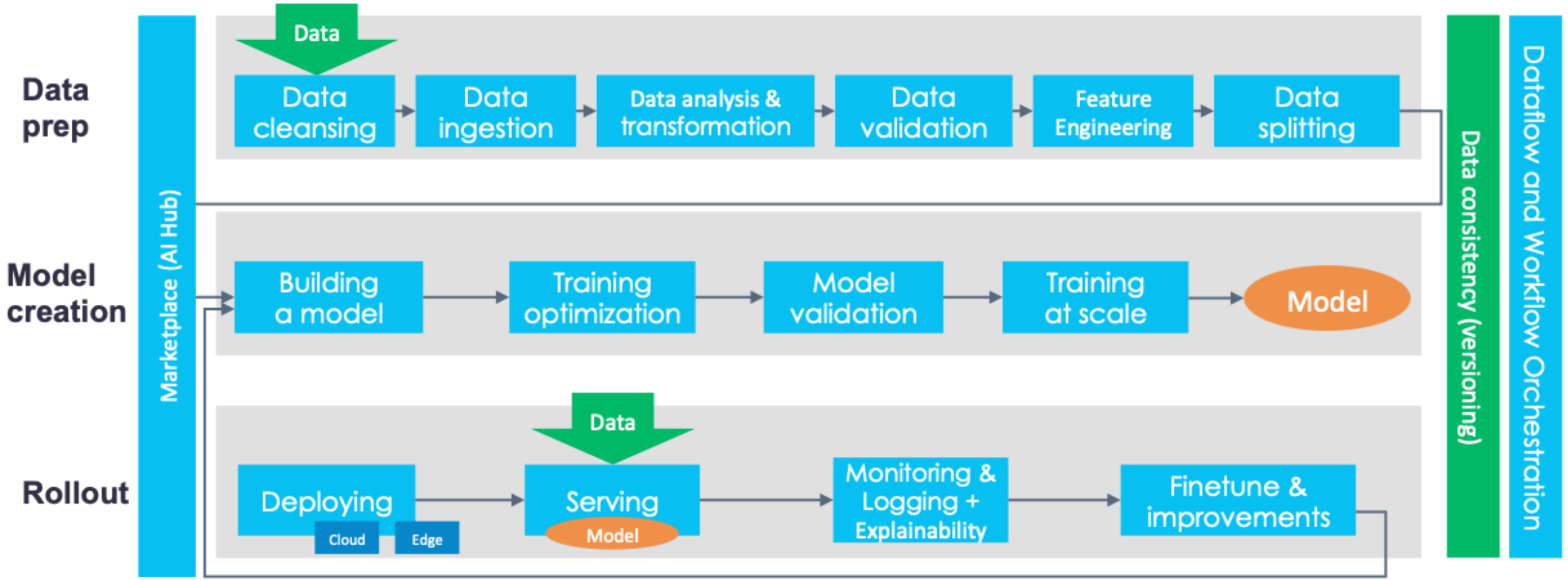
Machine learning platforms provide users with the tools necessary to develop, deploy, and improve machine learning, specifically machine learning algorithms.

Gartner defines a data science and machine learning platform as an **integrated** set of code-based libraries and low-code tooling that support the **independent use** by, and **collaboration** between, **data scientists** and their **business** and **IT** counterparts through all stages of the data science life cycle. These stages include **business understanding**, **data access and preparation**, **experimentation** and model creation, and **sharing** of insights. They also support machine learning engineering workflows including creation of data, feature, **deployment** and **testing pipelines**.

Source: https://www.gartner.com/

Machine learning platforms

2020s



Source: https://ianhellstrom.org/

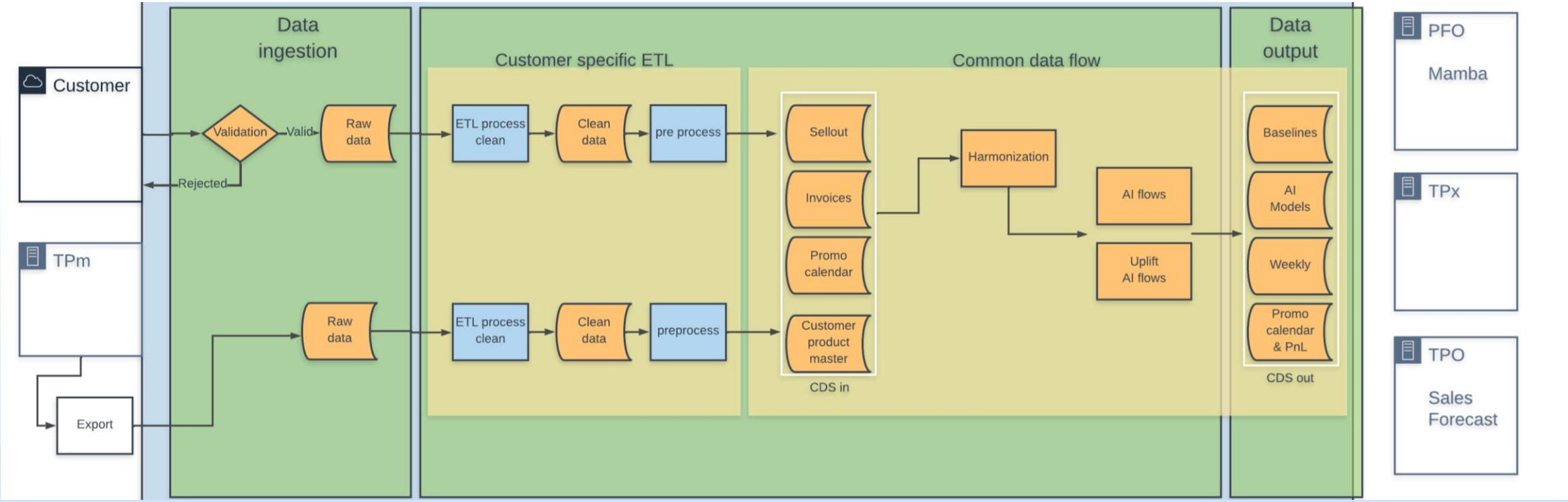
Our technology

Data flow

ETL	Harmonization	Analytics modules	Presentation
Validation	**Consistency**	**AI**	**Outputs**
Ingestion Data validation Filtering Read and clean TPM data Read and clean historical data Incremental data Concatenation	Featurization Data augmentation Filtering Labeling Flagging	PEA PFO Cannibalization TPO Power BI	Power BI TPO models Mamba models Baselines

Our technology

Data flow



Our technology

The ecosystem

Suricate

Autonomous data validation library

Hippo

Python framework to design data elaboration processes and Machine Learning experiments

Viper

Customizable API service

Mamba

RGM Analytics Module

Q&A

Thank you

https://www.xtel-group.com/

2024 Q2

Glauco Uri

glauco.uri@xtel-group.com

www.linkedin.com/in/glaucouri/

Ending with…

What will be the biggest challenge to face, what makes me most **uncertain** or **insecure**?