Data Warehouses meet Data Lakes

Mauro Pelucchi
Head of Global Data Science

Lightcast

PyCon LITHUANIA

Description

Many organizations have migrated their data warehouses to data lake solutions in recent years. With the convergence of the data warehouse and the data lake, a new data management paradigm has emerged that combines the best of both approaches: the bottom-up approach of big data and the top-down approach of a classic data warehouse.

In this talk, I will explain the current challenges of a data lake and how we can approach a modern data architecture with the help of PySpark, Apache Hudi, Delta Lake, or Iceberg. We will discuss how to organize data in a data lake to support real-time processing of applications and analysis across all types of data sets, both structured and unstructured. We will also examine how this approach provides the scalability needed to support enterprise-wide digital transformation and creates a single source of truth for multiple audiences.

I am a senior data scientist and big data engineer responsible for the designing of the "**Real-Time Labour Market Information System on Skill Requirements** " for eurostat CEDEFOP ETF Working together Learning for life European Training Foundation

Lightcast

UNIVERSITA' DEGLI STUDI DI MILANO BICOCCA

UNIVERSITAS STUDIORUM BERGOMENSIS

Currently, I work as the **Head of Global Data Science at Lightcast** , where I explore, design, and deliver innovative solutions related to labor market data.

I have collaborated with the **University of Milano-Bicocca** on various research projects related to labor market intelligence systems.

Additionally, I serve as a at the **University of Milano-Bicocca's Master Artificial Intelligence and Big Data Analytics** program and as a lecturer in **Computer Engineering at the University of Bergamo** .

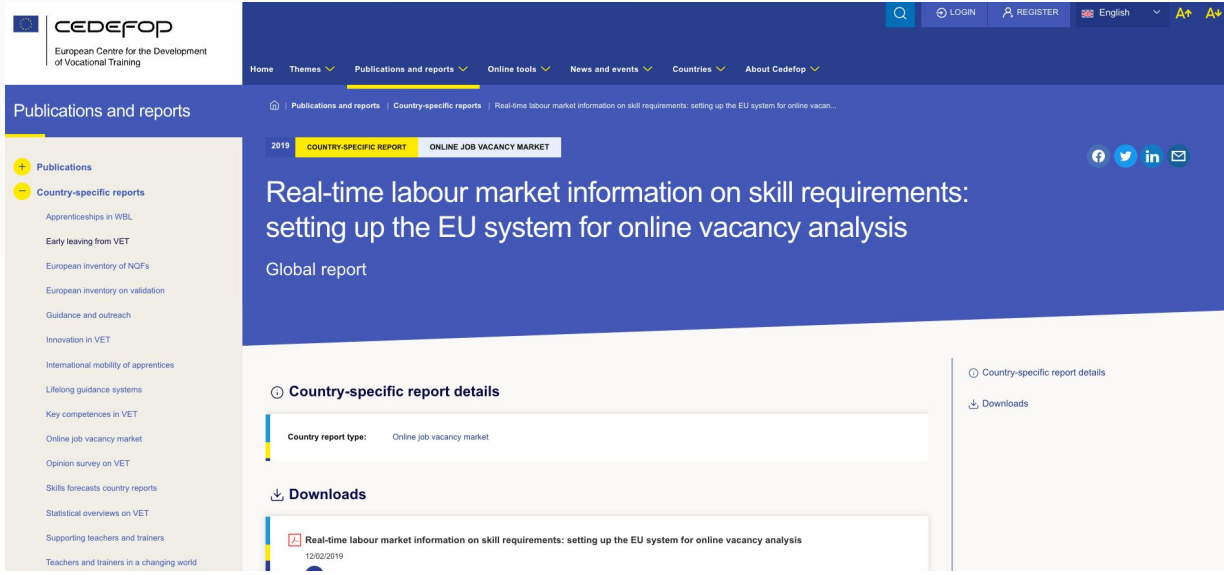
mauro.pelucchi@gmail.com
mauro.pelucchi@lightcast.io

https://github.com/mauropelucchi
https://x.com/mauropelucchi
https://www.linkedin.com/in/mauropelucchi/

https://github.com/mauropelucchi/pyconlt2025

Real-Time Labour Market Information System on Skill Requirements

Lightcast



Continuously evolving Labour Market

Context

Digitalization of professions
Relevance of Soft skills
Internationalization
New professions and skills emerging
Smart and Remote working
Impact of Covid-19 pandemic
Green transition
Artificial Intelligence

What we have / what we need

Official statistics

We already have official statistics, that are: **Representative**, **Strong** in terms of value

But we can benefit of additional, complementary information that could be: **fast**, to track what's happening now, **granular** and **fresh**, to capture emerging trends analyzing what companies are looking for

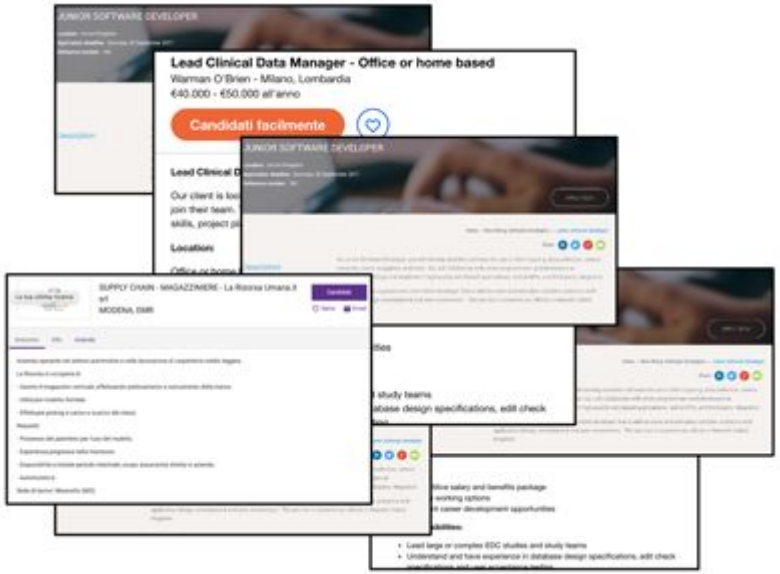
Big data

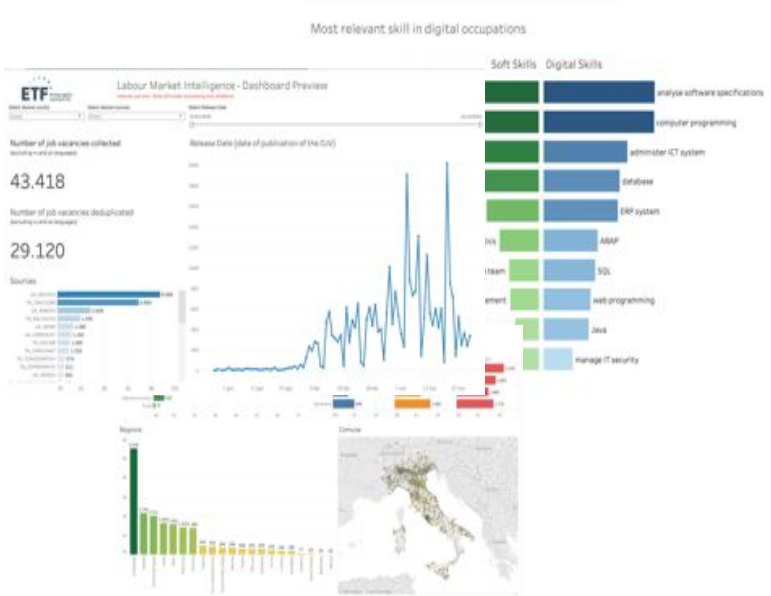
How to find a similar, complementary source of information?

Using Web Labour Market

Transform Online Job Advertisments…

…in insights and analytics





>1B

Downloaded pages.

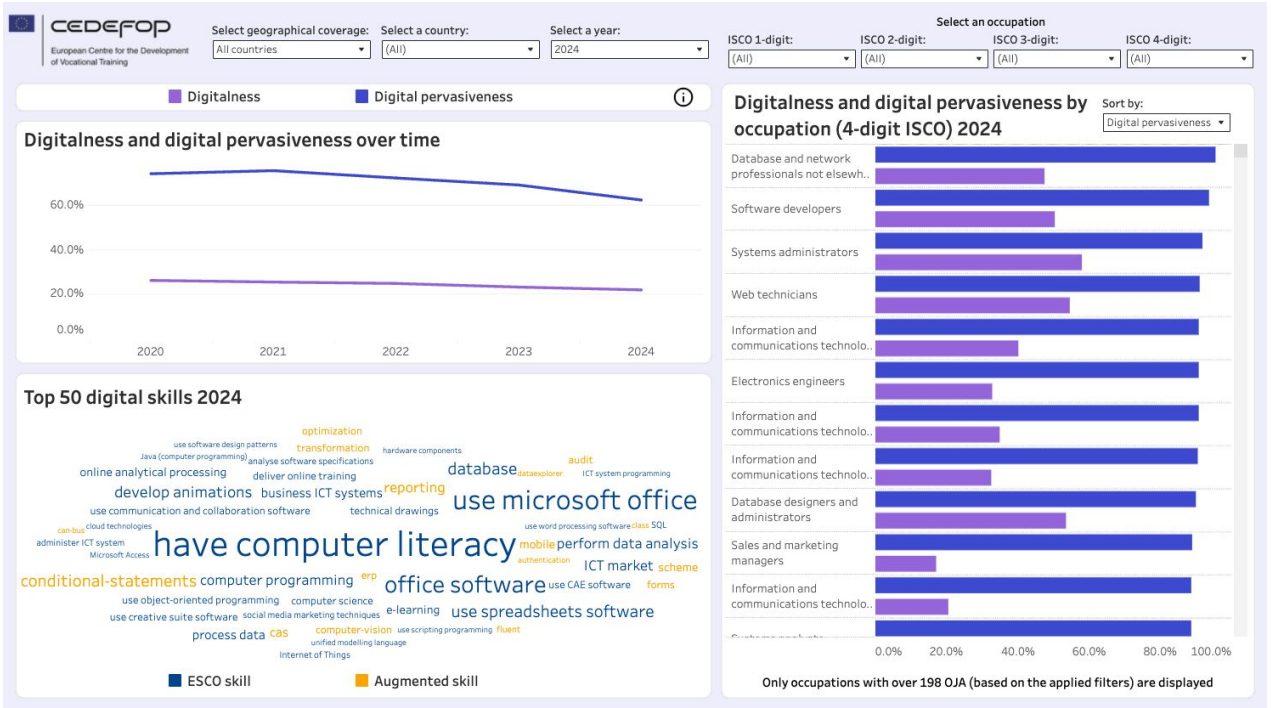
32

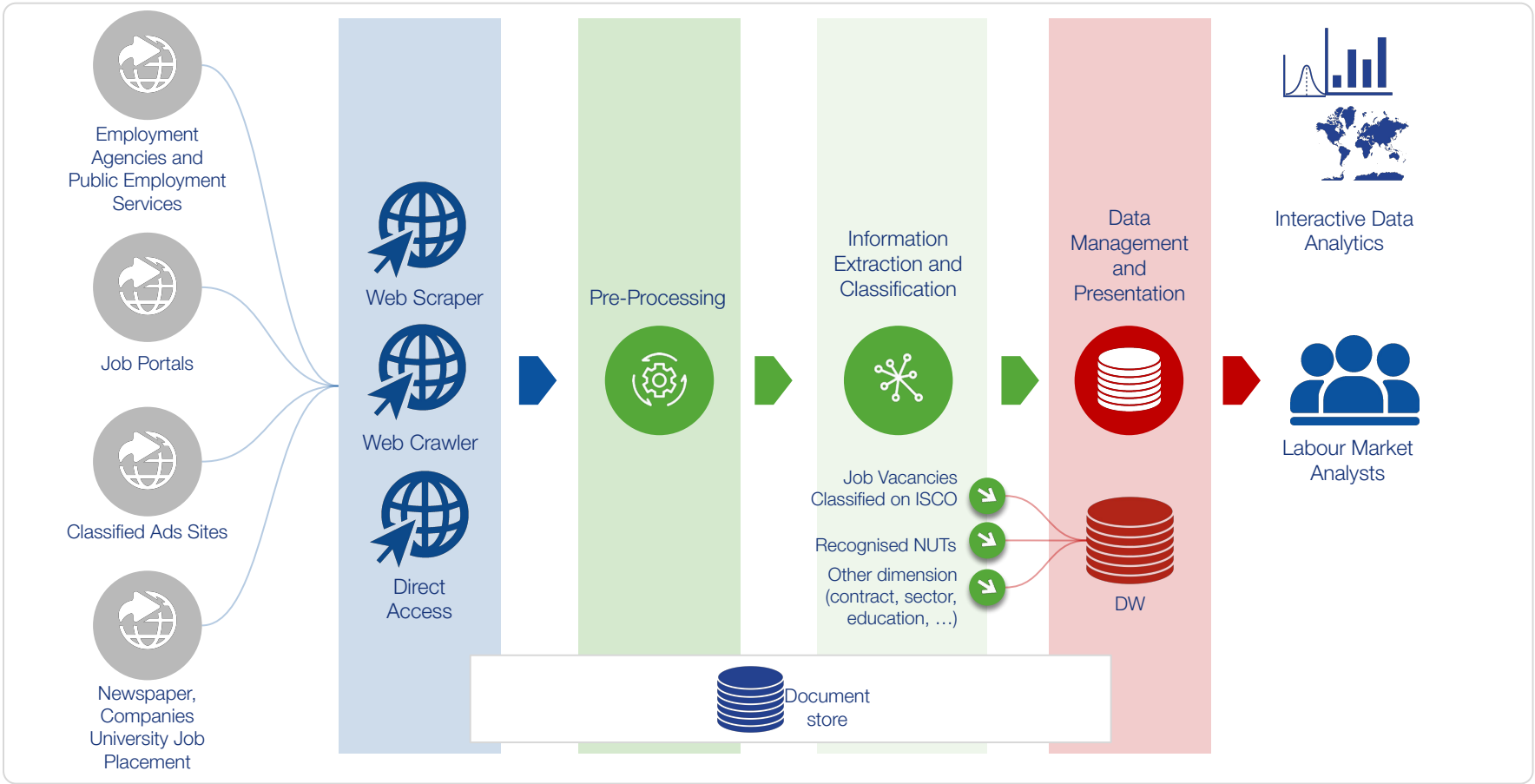
Countries.

9,645

Sources/Countries.

https://www.cedefop.europa.eu/en/tools/skills-online-vacancies




Employment Agencies and Public Employment Services
Job Portals
Classified Ads Sites
Newspaper, Companies University Job Placement
Web Scraper
Web Crawler
Direct Access
Pre-Processing
Information Extraction and Classification
Job Vacancies Classified on ISCO
Recognised NUTs
Other dimension (contract, sector, education, …)
Data Management and Presentation
DW
Interactive Data Analytics
Labour Market Analysts
Document store

JUNIOR DATA SCIENTIST & ANALYST PLACEMENT

London • Hybrid remote

Internship

You must create an Indeed account before continuing to the company website to apply

Apply on company site

As at Data Scientist a [redacted]ou will join the rapidly developing data team, who are responsible for measurement solutions and modelling expertise help a diverse client-set understand the true value of their media investment, create compelling data stories on how to drive growth, and automate the insights into the planning cycle through their advance and integrated tech stack.

We are looking fo inquisitive, articulate, numerate and above-all, enthusiastic placement students to support the wider team in delivering these critical insights and building on the capabilities of our product.

You will be part of a close-knit and friendly team who share results and celebrate success together.

[redacted]s a media agency that's made differently. We're purpose-led, data-driven and proudly independent. Our independence means we can focus 100% on doing the right thing to secure success for our clients and our brilliant people. We are trusted to deliver that success for some of the UK's most ambitious and complex organisations, including SunLife, Guide Dogs, RNLI, Laithwaites and National Trust.

WHAT YOU'LL BE DOING

Work closely with business to identify issues that can be resolved using data solutions effectively for decision making

Machine learning tools and statistical methods to solve complex problems

Build algorithms and design experiments to merge, manage, interrogate, and extract data to supply tailored reports to colleagues, clients and wider areas in company

Support the account management and planning teams across all facets of campaign measurement across media channels

Develop automated data processes using Python/R

Ability to organise a variety of large data sets

Undertake regular analysis and reporting for retained clients

Maintain clear and coherent communication, both verbal and written, to understand data needs and report results

Working with the Datalab team and other senior business stakeholders to develop analytical propositions

THE SKILLS YOU WILL BRING

Highly numerate undergraduate studying a relevant degree in mathematics, statistics, econometrics or computer science

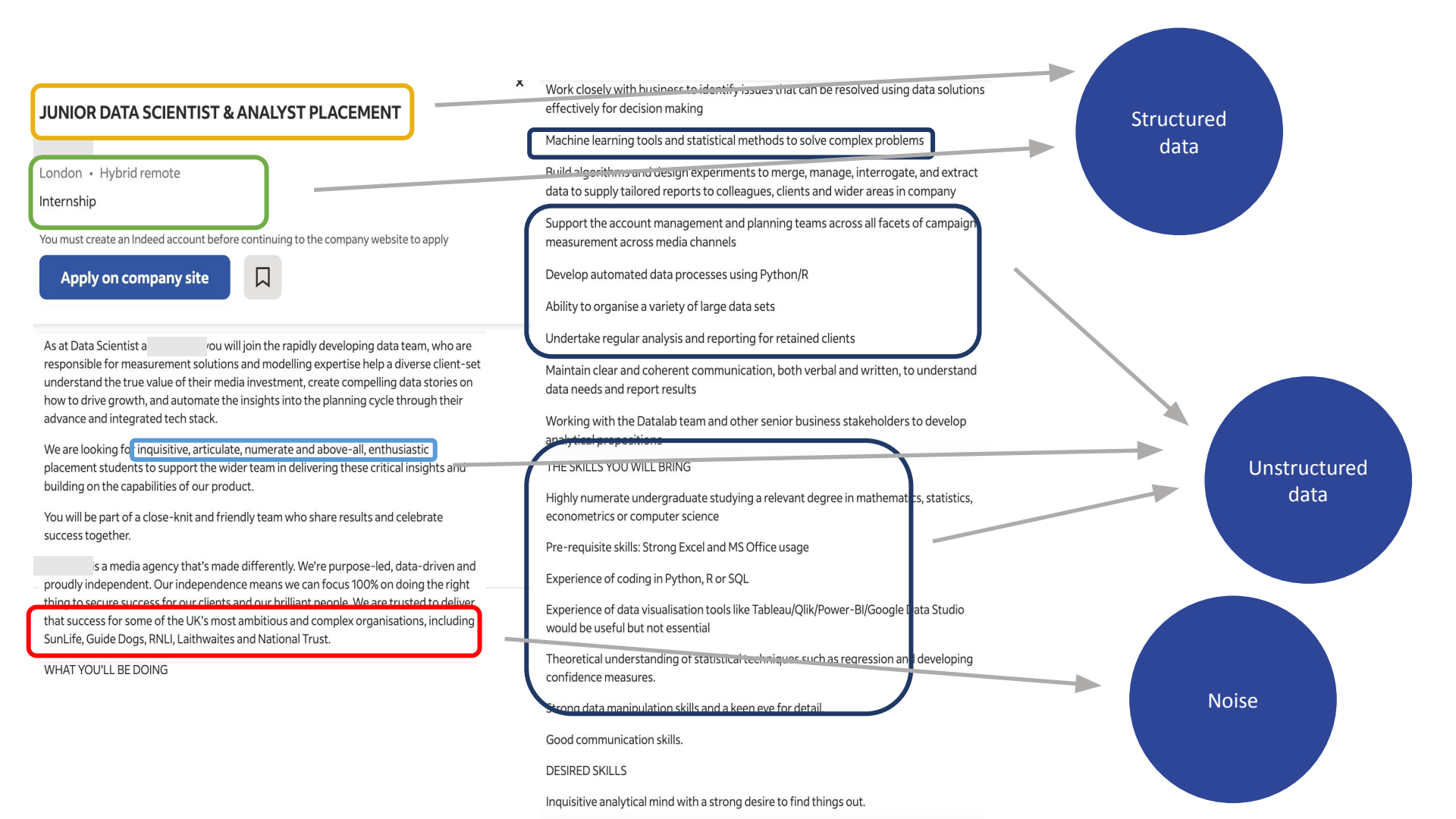
Pre-requisite skills: Strong Excel and MS Office usage

Experience of coding in Python, R or SQL

Experience of data visualisation tools like Tableau/Qlik/Power-BI/Google Data Studio would be useful but not essential

Theoretical understanding of statistical techniques such as regression and developing confidence measures.

Strong data manipulation skills and a keen eye for detail.


JUNIOR DATA SCIENTIST & ANALYST PLACEMENT
London • Hybrid remote
Internship
You must create an Indeed account before continuing to the company website to apply
Apply on company site
As at Data Scientist a you will join the rapidly developing data team, who are responsible for measurement solutions and modelling expertise help a diverse client-set understand the true value of their media investment, create compelling data stories on how to drive growth, and automate the insights into the planning cycle through their advance and integrated tech stack.
We are looking fo inquisitive, articulate, numerate and above-all, enthusiastic placement students to support the wider team in delivering these critical insights and building on the capabilities of our product.
You will be part of a close-knit and friendly team who share results and celebrate success together.
s a media agency that's made differently. We're purpose-led, data-driven and proudly independent. Our independence means we can focus 100% on doing the right thing to secure success for our clients and our brilliant people. We are trusted to deliver that success for some of the UK's most ambitious and complex organisations, including SunLife, Guide Dogs, RNLI, Laithwaites and National Trust.
WHAT YOU'LL BE DOING
Work closely with business to identify issues that can be resolved using data solutions effectively for decision making
Machine learning tools and statistical methods to solve complex problems
Build algorithms and design experiments to merge, manage, interrogate, and extract data to supply tailored reports to colleagues, clients and wider areas in company
Support the account management and planning teams across all facets of campaign measurement across media channels
Develop automated data processes using Python/R
Ability to organise a variety of large data sets
Undertake regular analysis and reporting for retained clients
Maintain clear and coherent communication, both verbal and written, to understand data needs and report results
Working with the Datalab team and other senior business stakeholders to develop analytical propositions
THE SKILLS YOU WILL BRING
Highly numerate undergraduate studying a relevant degree in mathematics, statistics, econometrics or computer science
Pre-requisite skills: Strong Excel and MS Office usage
Experience of coding in Python, R or SQL
Experience of data visualisation tools like Tableau/Qlik/Power-BI/Google Data Studio would be useful but not essential
Theoretical understanding of statistical techniques such as regression and developing confidence measures.
Strong data manipulation skills and a keen eye for detail
Good communication skills.
DESIRED SKILLS
Inquisitive analytical mind with a strong desire to find things out.
Structured data
Unstructured data
Noise

Challenges

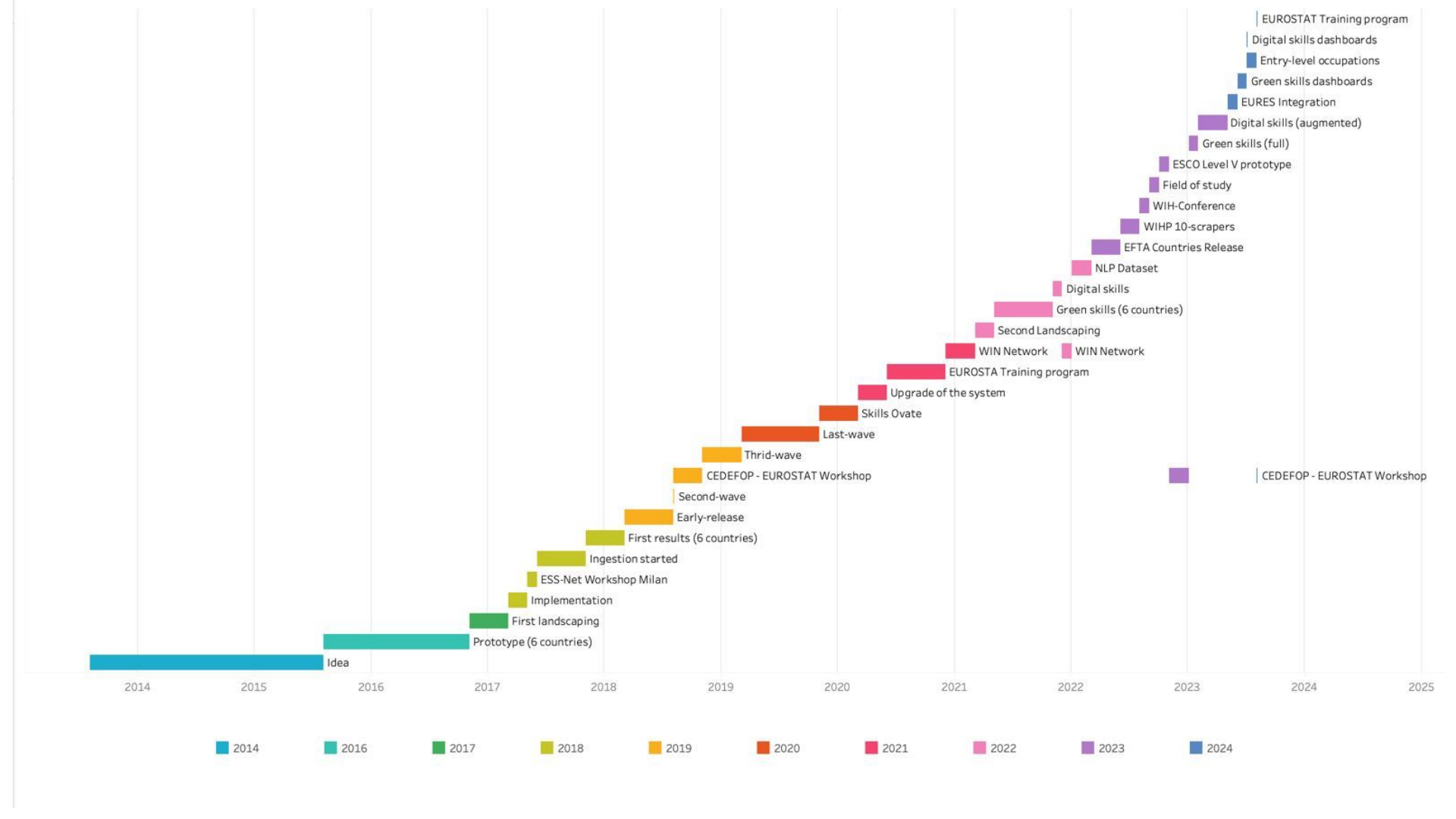
Handle a huge amount of near real time data

Data coming from web > Need to detect and reduce noise

Multi language environment

Need to relate to classification standards

Find a way to summarize and present a wide and complex scenario


EUROSTAT Training program
Digital skills dashboards
Entry-level occupations
Green skills dashboards
EURES Integration
Digital skills (augmented)
Green skills (full)
ESCO Level V prototype
Field of study
WIH-Conference
WIHP 10-scrapers
EFTA Countries Release
NLP Dataset
Digital skills
Green skills (6 countries)
Second Landscaping
WIN Network
WIN Network
EUROSTA Training program
Upgrade of the system
Skills Ovate
Last-wave
Thrid-wave
CEDEFOP - EUROSTAT Workshop
CEDEFOP - EUROSTAT Workshop
Second-wave
Early-release
First results (6 countries)
Ingestion started
ESS-Net Workshop Milan
Implementation
First landscaping
Prototype (6 countries)
Idea
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2014
2016
2017
2018
2019
2020
2021
2022
2023
2024

Topics

- Datawarehouse and business intelligence processes
- Top down approach and architecture for traditional BI systems
- What is the question?
- Bottom-up approach
- A new way or a different way?
- Delta lake
- Technologies: delta.io, hudi and iceberg
- The next steps

@Source: Cukier – 2014 TEDx

What is America's favourite type of cake?

@Source: Cukier – 2014 TEDx

Apple Pie

How do we know?

Retail sales data

ETL and data warehouse

Provisioning to our marketing and sales team

What is the type of cake with the highest turnover?

How do we know?
Business intelligence process

Collecting of the data

Retail sales data

Optimizing for analysis

ETL and data warehouse

Provisioning

Provisioning to our marketing and sales team

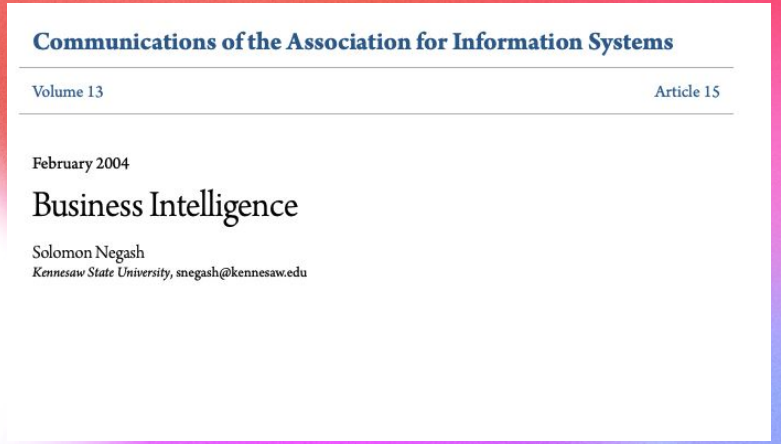
Analytics

What is the type of cake with the highest turnover?

Business Intelligence

Business Intelligence is an initiative in companies in order to collect, transform and provide data for users to plan, control and steer the company to achieve the company's goals

Negash, Solomon, and Paul Gray. "Business intelligence." *Handbook on decision support systems 2*. Springer, Berlin, Heidelberg, 2008. 175-193.



BI Architecture

Question

What is America's favourite type of cake?

Source systems

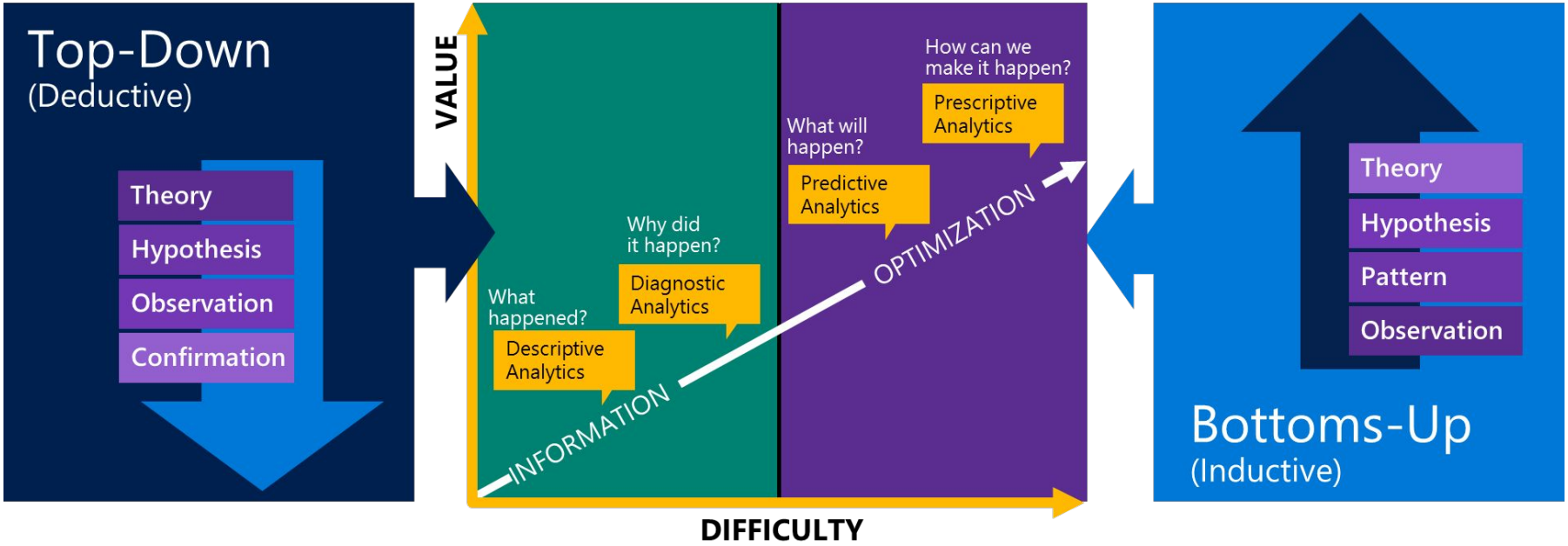
Collecting & transform

Storage

Provisioning

Application

Top-Down approach



Credits: James Serra @ Microsoft

What was the question?

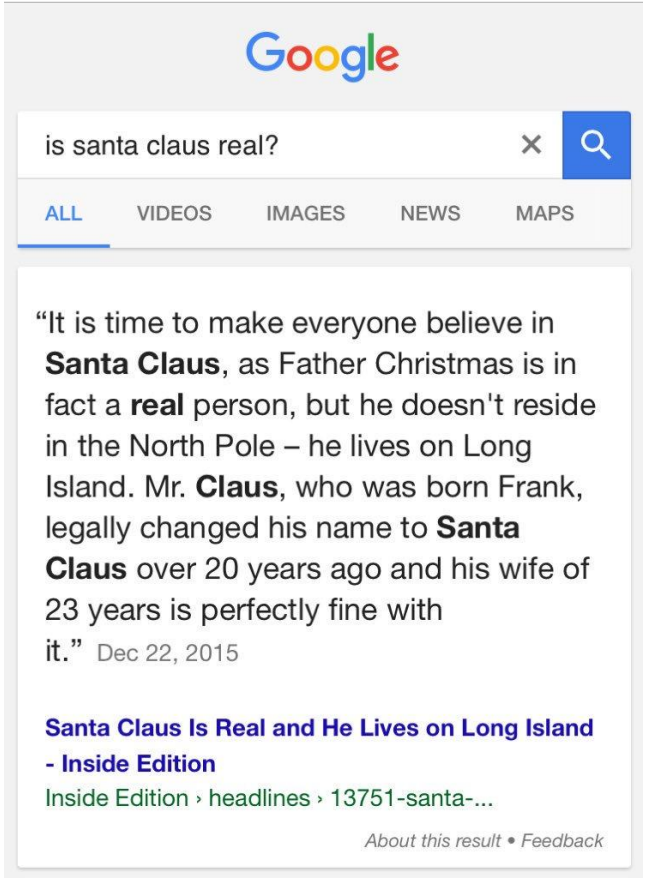
What is America's
favourite type of cake?

What was the question?

What is America's
favourite type of cake?

Favourite vs
Retail sales data

New questions



New sources

airbnb

Tide
amazon dash BUTTON

TESLA

flightradar24 LIVE AIR TRAFFIC

planet.

PLANET OS

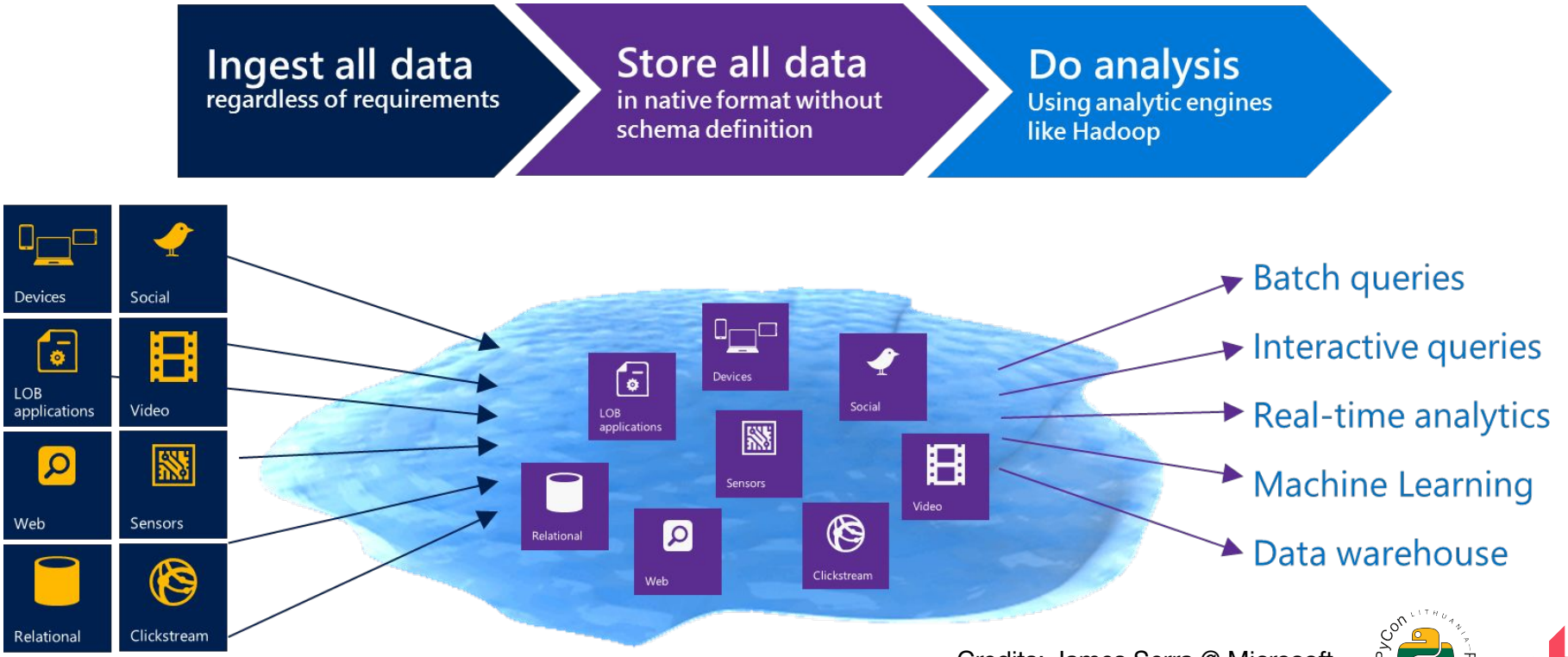
Propeller

indeed

Terra Bella

myfitnesspal
Calorie Counter & Diet Tracker

Bottom-up approach



Data Warehouses meet Data Lakes

Credits: James Serra @ Microsoft

Problem statement

What we need?

- Automatic “Add data in minutes”
- From Reporting to ML “User configurable”
- Any data, any sources
- Scaling, support exponential growth
- Accessibility ”tools and log-on”

Data warehouse

Data warehouse: collection of data (it is a database)
that comes from other databases.

- Integrated: contains data on different but related topics
- Subject Oriented: oriented according to specific topics of analysis but not to applications
- Time Variant: the time horizon of the data in the DWH is usually longer or different than the horizon of the data managed by the management system
- Non-volatile: data in a DWH, once loaded, are not modifiable

Data lake

- A data lake is a centralised repository for storing all structured and unstructured data at any scale.

- Data can be stored as is, without having to structure it first, and different types of data analysis can be performed – from control panels and visualisations to Big Data processing, real-time data analysis and machine learning to make better decisions.

The data lake paradigm

Unlock the value of the data earlier

Machine learning ready

Data (raw) access

Model data, schema free

Rapid change

Data lake vs data warehouse

Data Warehouse

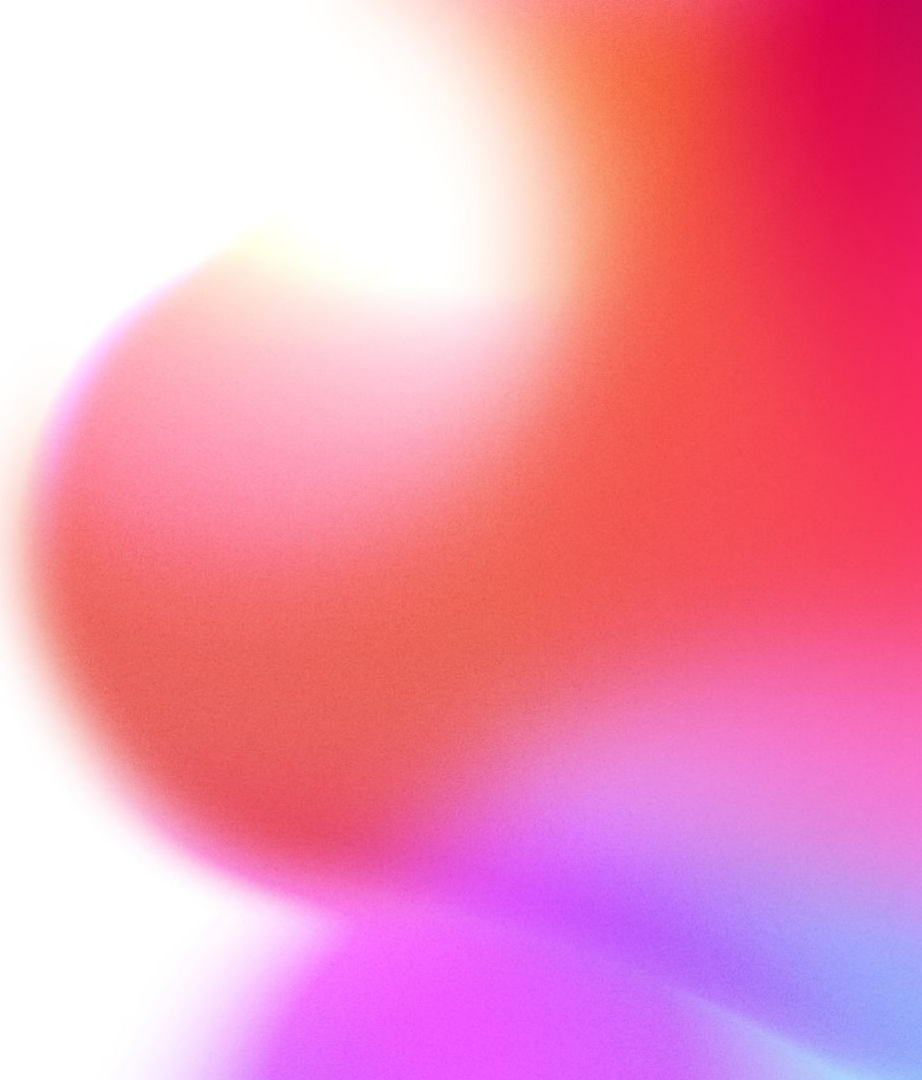
- Aggregated Subsets
- On-Demand Views
- Curated By Experts
- Structured - Tables, Views, Reports. Limited Context
- Data Quality Is Known And Tracked

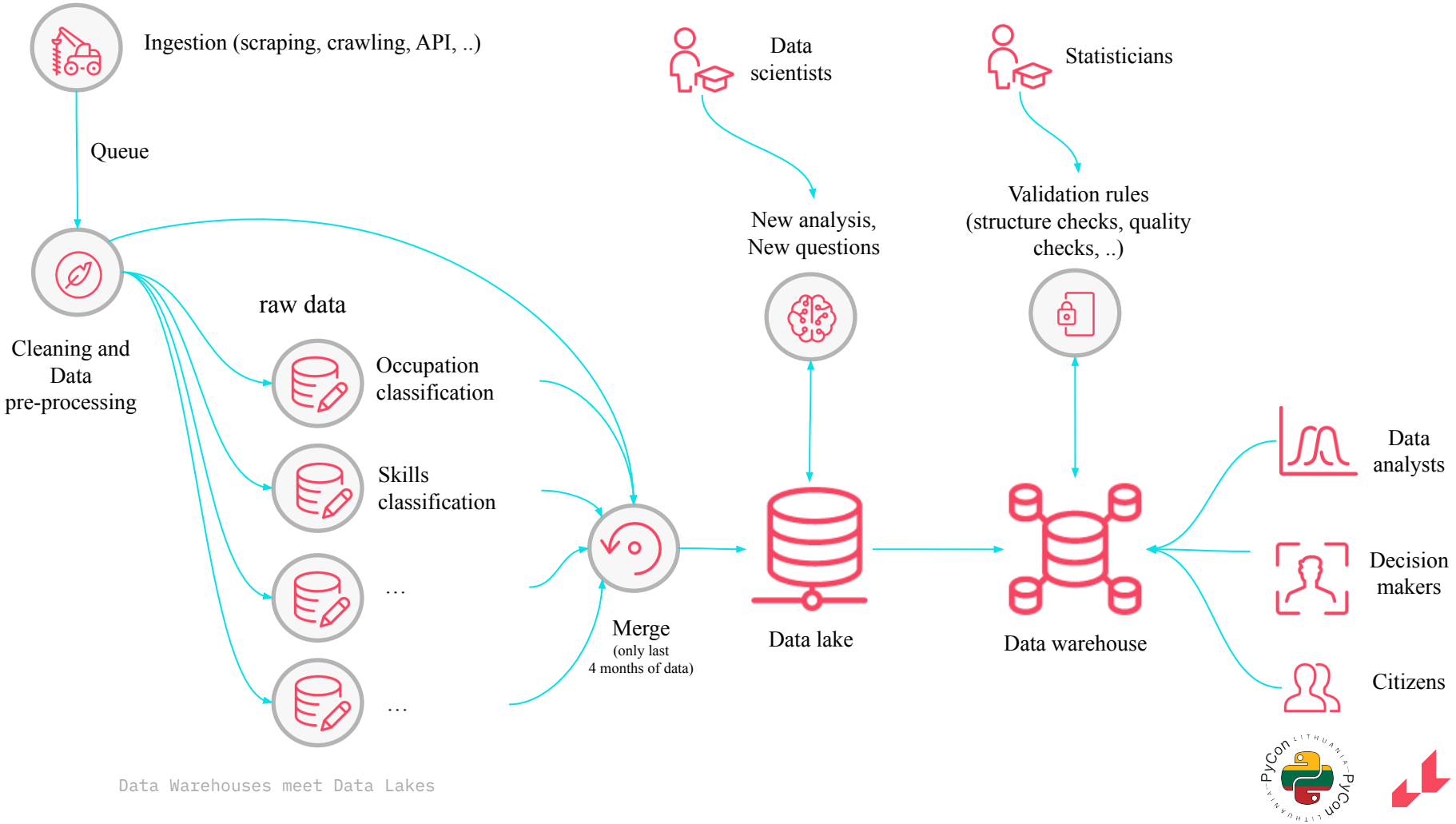
Data Lake

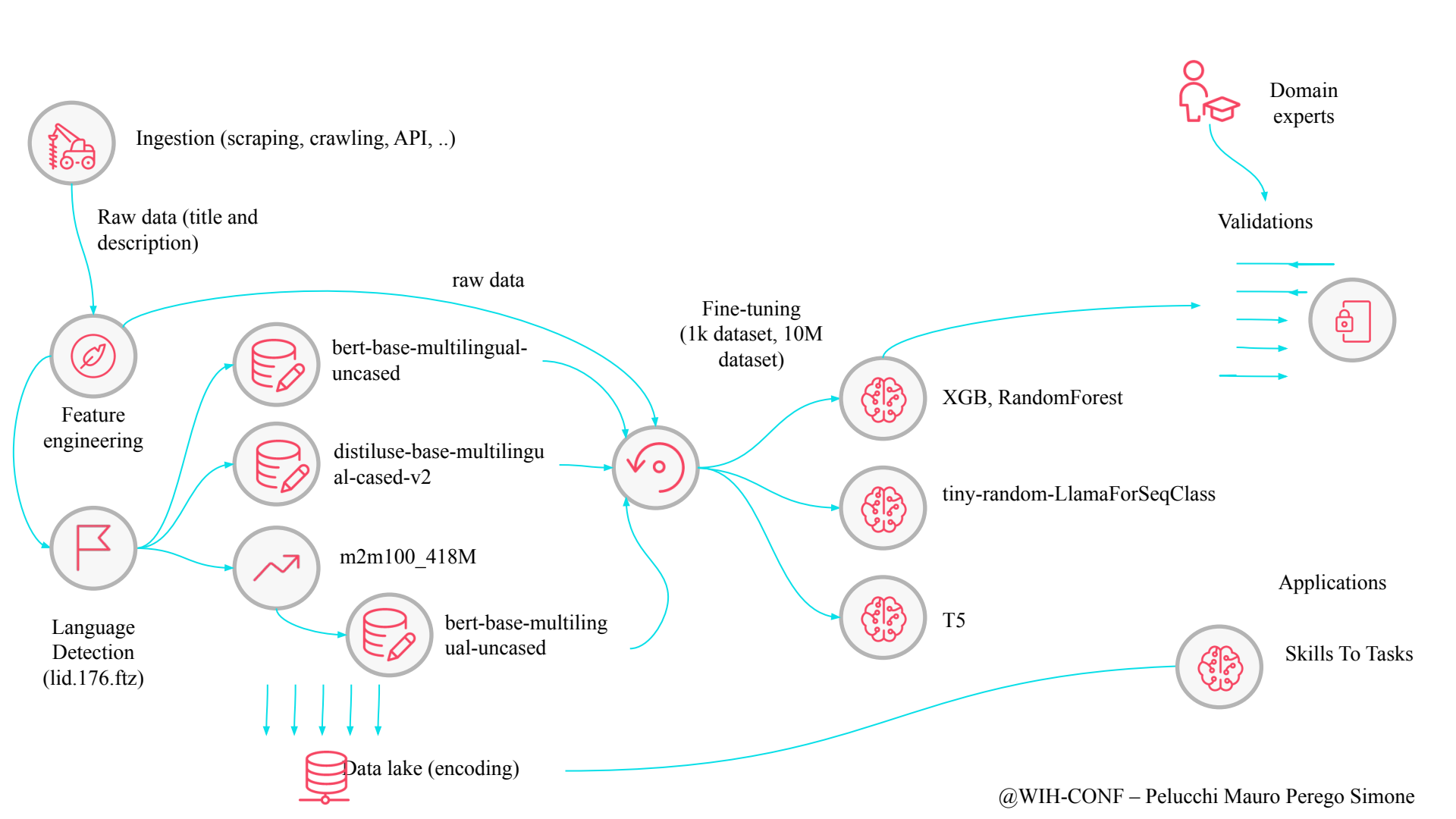
- Store Everything As-is
- Let Business Decide What They Need
- Support Rapid Change
- Provide Data Lineage and History Tracking and Visualization
- Unstructured - Key-Word Search
- Data Is Available In Various States from Raw to Fully Conformed
- Quality Metrics Often Not Available

Data Lake challenge

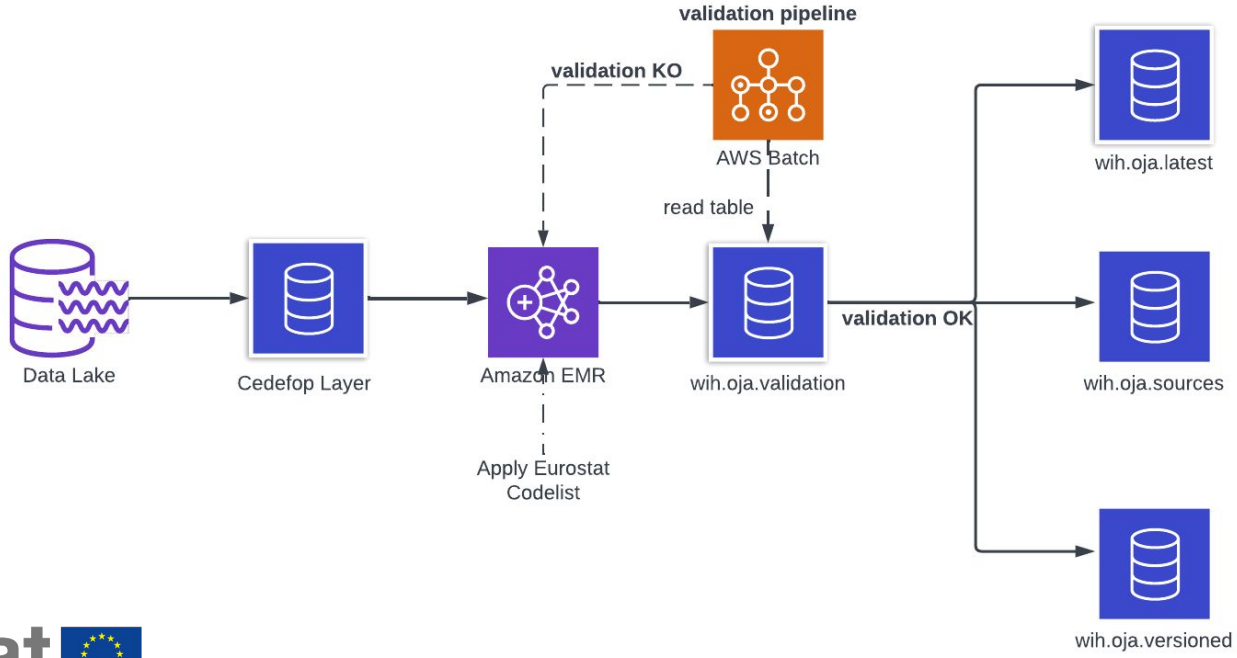
- Complex, multi- layered architecture
- Unknown storage & scalability
- Working with un-curated data
- Performance
- Data retrieval
- Data quality
- Governance
- Redundant effort




Ingestion (scraping, crawling, API, ..)
Queue
Cleaning and
Data
pre-processing
raw data
Occupation
classification
Skills
classification
…
…
Merge
(only last
4 months of data)
Data
scientists
New analysis,
New questions
Data lake
Statisticians
Validation rules
(structure checks, quality
checks, ..)
Data warehouse
Data
analysts
Decision
makers
Citizens


Ingestion (scraping, crawling, API, ..)
Raw data (title and description)
raw data
Feature engineering
Language Detection (lid.176.ftz)
bert-base-multilingual-uncased
distiluse-base-multilingual-cased-v2
m2m100_418M
bert-base-multilingual-uncased
Data lake (encoding)
Fine-tuning (1k dataset, 10M dataset)
XGB, RandomForest
tiny-random-LlamaForSeqClass
T5
Domain experts
Validations
Applications
Skills To Tasks

Data Pipeline



eurostat

Concepts
Columnar Data Formats

Filters

Filters are not the only “predicate” that can be pushed down

Query push down

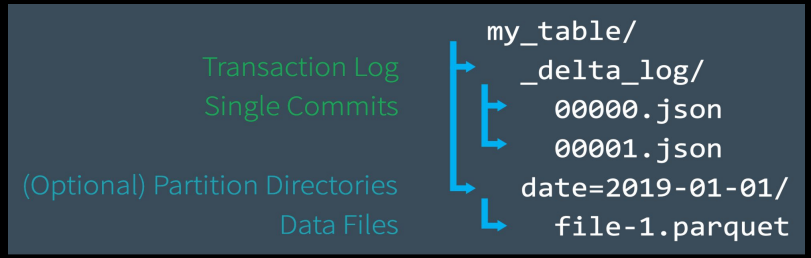
Column selection can also be pushed down
- With a database like PostgreSQL, this is done with a SELECT statement
- For files, we require a Columnar File Format

Storage

Data is stored by column, not by row
- Parquet and ORC
- Delta lake format: Delta.io, Hudi, Iceberg

What is Delta.io?

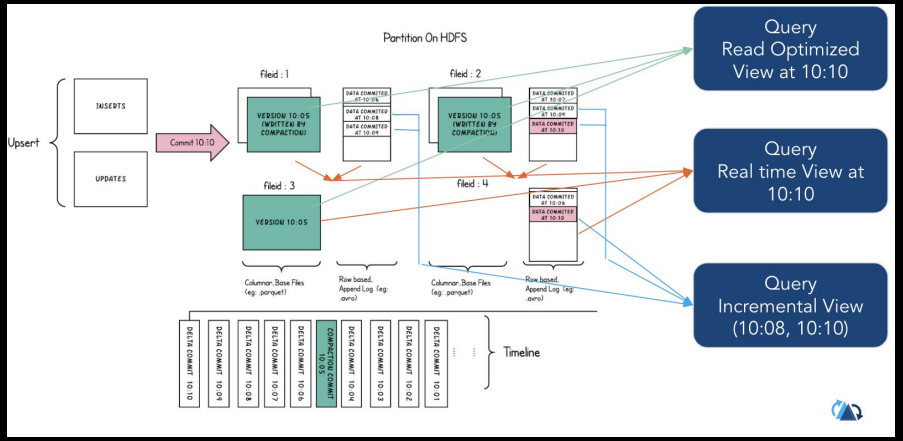
- Technology designed to be used with Apache Spark to build robust data lakes -> Has strong Spark integration, especially with Databricks tools.
- ACID transactions on Spark
- Scalable metadata handling
- Upserts and deletes
- Fully configurable/optimizable
- Structured streaming support
- Based on transaction logs (JSON) appended per commit
- Good for full-stack Lakehouse scenarios: ML, BI, streaming, and batch.



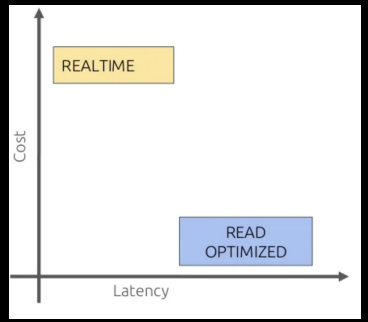
- Best if you're heavily invested in Spark and want tight Databricks integration

What is Apache Hudi?

- Schema evolution
- Emphasizes incremental data ingestion,with copy-on-write or merge-on-read modes.
- Tracks commit history via timeline
- Compatibility of Spark and Kafka
- Query Hudi with Athena, Integrated in AWS



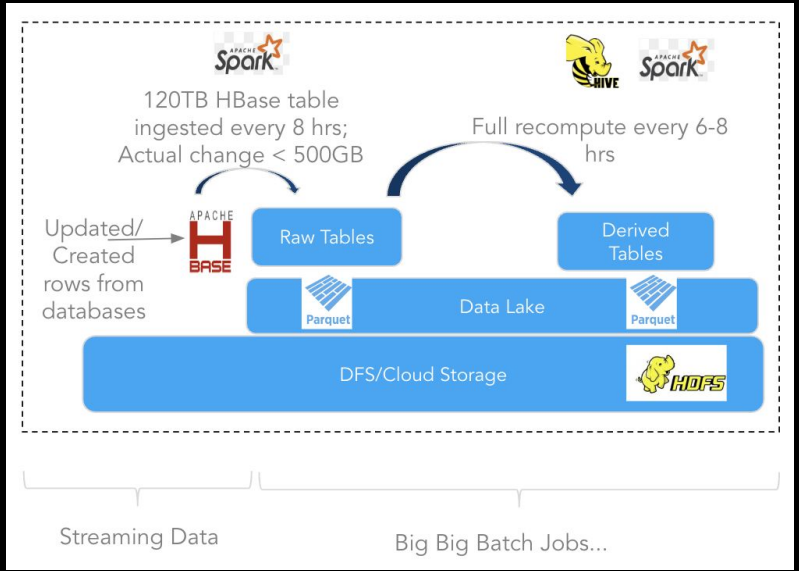
- Designed around fast updates, upserts, and streaming pipelines
- Good for data lakes with change-heavy workloads (e.g., transactional systems)



@Source:
Apache Hudi

Data Lake In-efficiency

- Batch ingestion too slow
- Rewrite entire tables/partitions several times
- ETLs off raw data have no smarts to recompute
- Late arriving data is a nightmare



@Source: Apache Hudi

What is Iceberg?

- Scalable format for tables. Table format optimized for immutability, with manifest files tracking data/metadata.
- Treats table as a snapshot, allowing fast time travel and rollback
- No dependency from Spark
- Robust schema and partitioning changes
- Fast query planning

- Suited for high concurrency, big datasets, and cloud-native warehouses

```
val df = spark.read
         .format("iceberg")
         .load("s3://datalake/d_manufacturers")
```

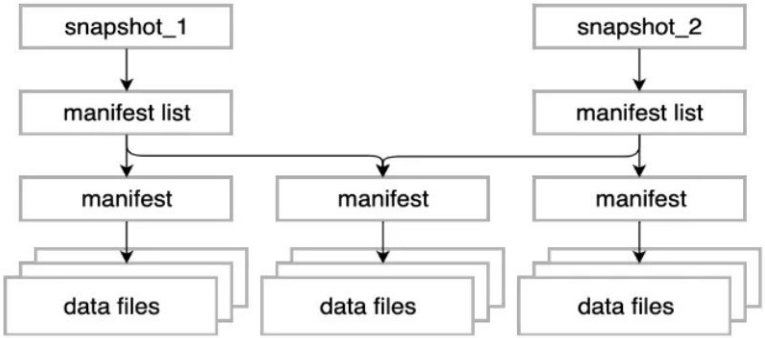


	Table Metadata	Transaction Atomicity	Isolation Levels
Delta Lake	Transaction Log + Metadata Checkpoints	Atomic Log Appends	Serializability, Strict Serializability
Hudi	Transaction Log + Metadata Table	Table-Level Lock	Snapshot Isolation
Iceberg	Hierarchical Files	Table-Level Lock	Snapshot Isolation, Serializability

Table 1: Lakehouse system design features: How they store table metadata, how they provide atomicity for transactions, and what transaction isolation levels they provide.

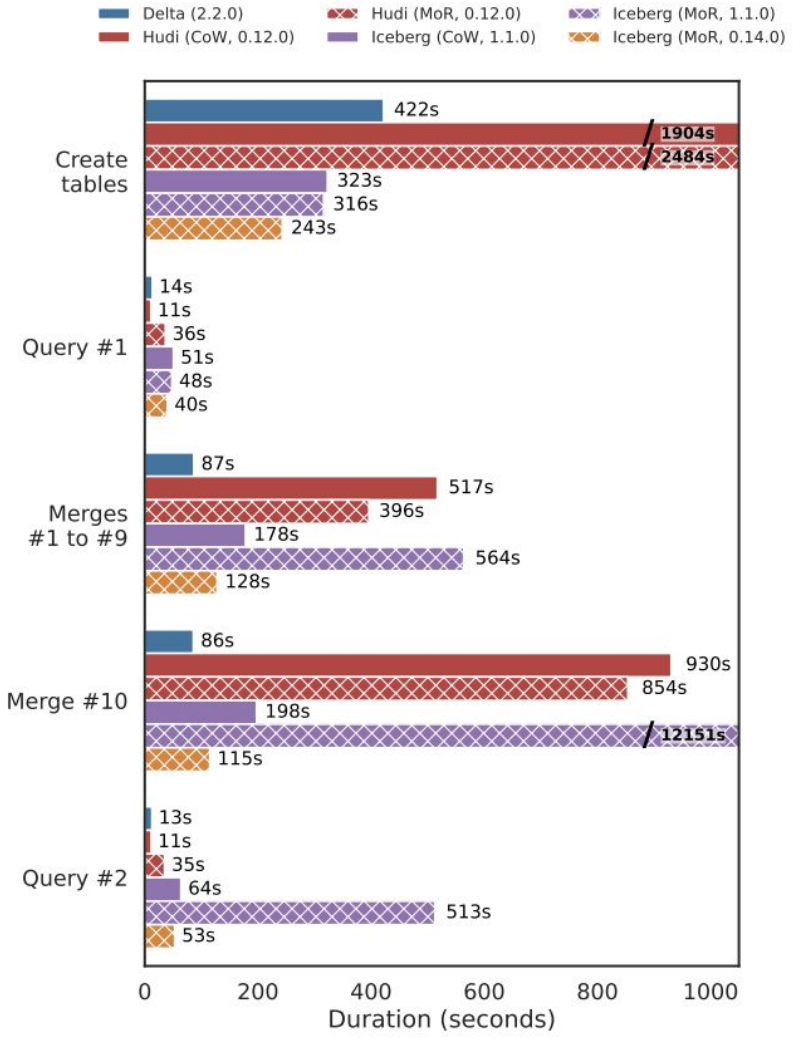


Figure 2: Performance of the 100 GB TPC-DS incremental refresh benchmark. The benchmark loads data, runs five queries (Q3, Q9, Q34, Q42, and Q59), then merges changes into the table ten times and runs these queries again. Hudi ran compaction in merge iteration 10, so we report it separately from 1–9. Iceberg results were run with a higher S3 connection pool size due to timeout errors [16]. Delta and Hudi results were run with the default EMR configuration.

Jain, P., Kraft, P., Power, C., Das, T., Stoica, I., & Zaharia, M. (2023). Analyzing and Comparing Lakehouse Storage Systems. In *Conference on Innovative Data Systems Research (CIDR'23). https://www. cidrdb. org/cidr2023/papers/p92-jain. pdf*.

Feature	Apache Iceberg	Apache Hudi	Delta Lake
Origin	**Netflix**, now **Apache project**	**Uber**, now **Apache project**	**Databricks**
ACID Transactions	Yes (via manifest + metadata layers)	Yes (via timeline + commit logs)	Yes (via transaction log)
Time Travel	Yes	Yes	Yes
Schema Evolution	Full (add, drop, rename columns)	Limited (rename not fully supported)	Full (add, drop, rename columns)
Partition Evolution	Yes	No	No
Compaction	Not required	Required (for merge-on-read tables)	Optional
Streaming Support	Strong (Spark, Flink, Kafka, etc.)	Strong (focus on near real-time ingest)	Strong (Databricks, Spark Streaming)
Query Engines	Spark, Flink, Trino, Presto, Dremio	Spark, Flink, Hive, Presto	Spark, Presto, Trino (best with Databricks)
Cloud Native Design	Yes	Not fully (still Hadoop-centric)	Partially (designed for Delta Lakehouse)
Metadata Management	Centralized metadata via manifest	Timeline-based with commit logs	JSON transaction logs
Indexing	Planned (some custom impls exist)	Bloom & column stats index	No native indexing
Data Layout Optimization	Hidden partitioning, metadata pruning	File sizing + clustering	Z-ordering (Databricks)
GitHub Commits	12,800+	9,500+	6,200+
PRs Merged	12,000+	8,800+	5,700+
Open PRs	300+	400+	250+

Use Case	Best Choice	Notes
Large-scale batch analytics	Iceberg	Best metadata performance and scalability
Real-time ingestion + fast updates	Hudi	Optimized for change data capture (CDC), upserts
Spark + Delta ecosystem users	Delta Lake (esp. on Databricks)	Seamless integration with ML & BI tooling

Choose **Iceberg** if your priority is scalable, cloud-native batch analytics with complex metadata operations.

Choose **Hudi** if you need near real-time data lakes with fast ingestion and frequent updates.

Choose **Delta Lake** if you're in a Spark-centric environment and want simplicity and native support.

Thanks

Mauro.Pelucchi@gmail.com

Mauro.Pelucchi@lightcast.io