Data Warehouses meet Data Lakes

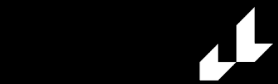
Mauro Pelucchi
Head of Global Data Science

Lightcast

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

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

Data Warehouses meet Data Lakes

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

UNIVERSITAS STUDIORUM BERGOMENSIS

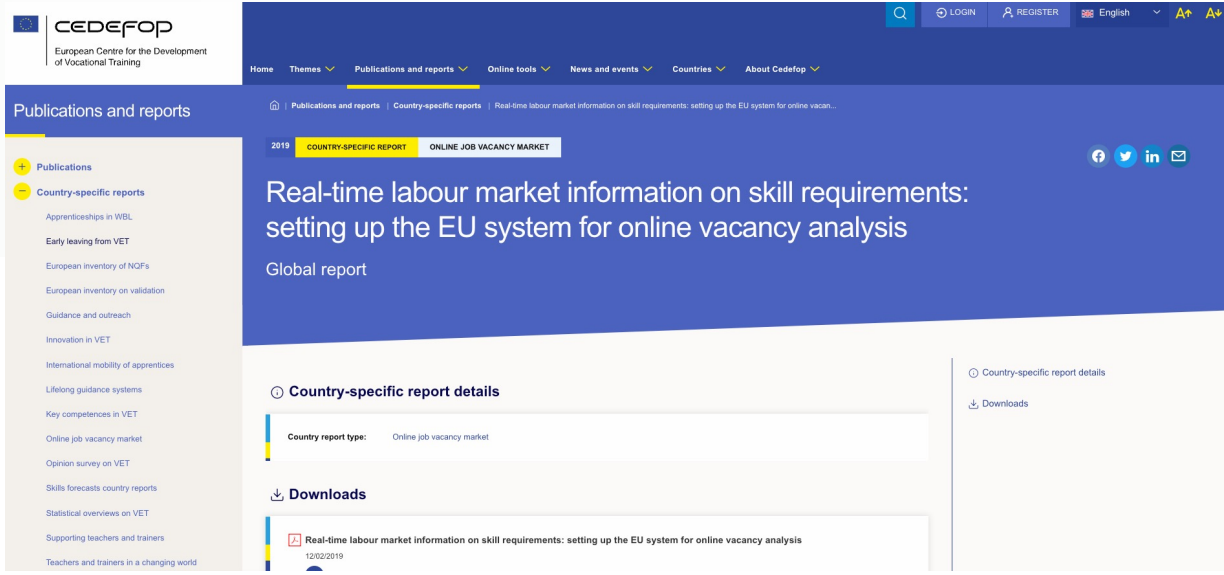
mauro.pelucchi@gmail.com
mauro.pelucchi@lightcast.io

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

https://github.com/mauropelucchi/pyconit2023

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

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

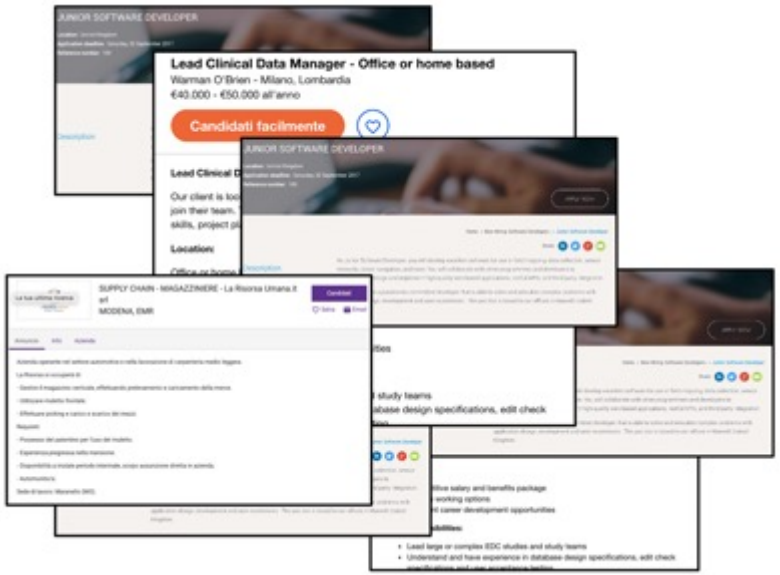
Big data

How to find a similar, complementary source of information?

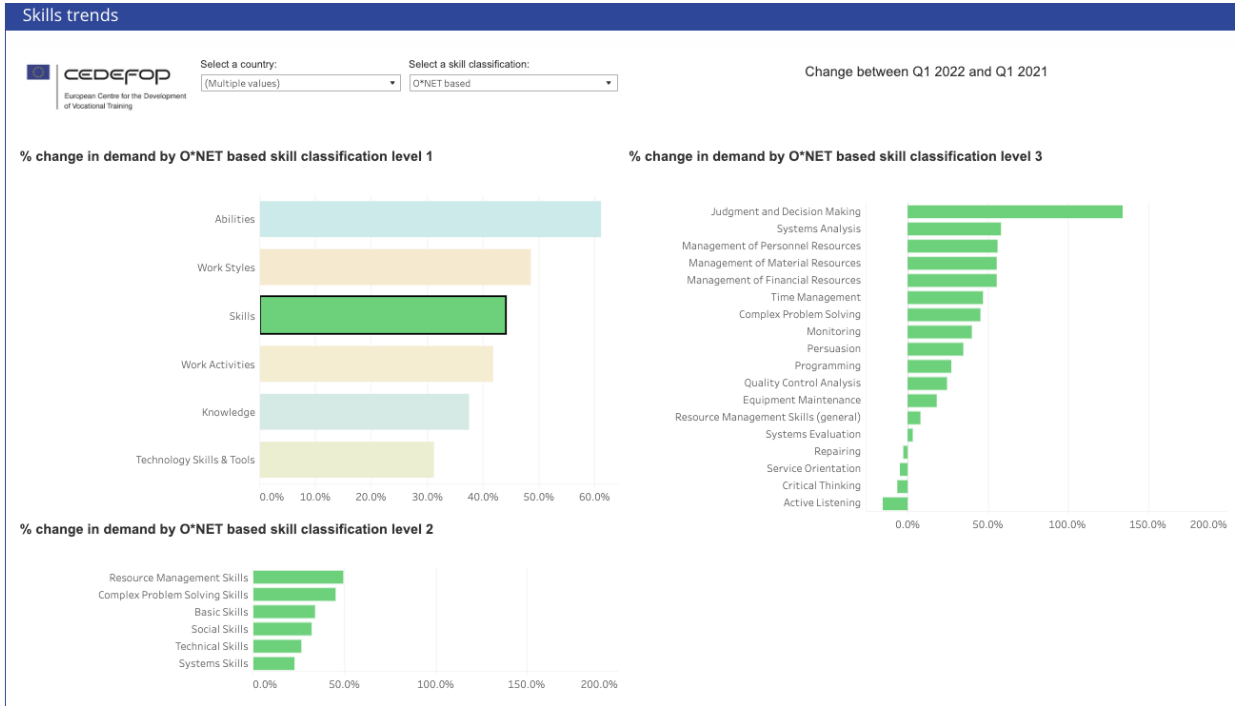
Using Web Labour Market

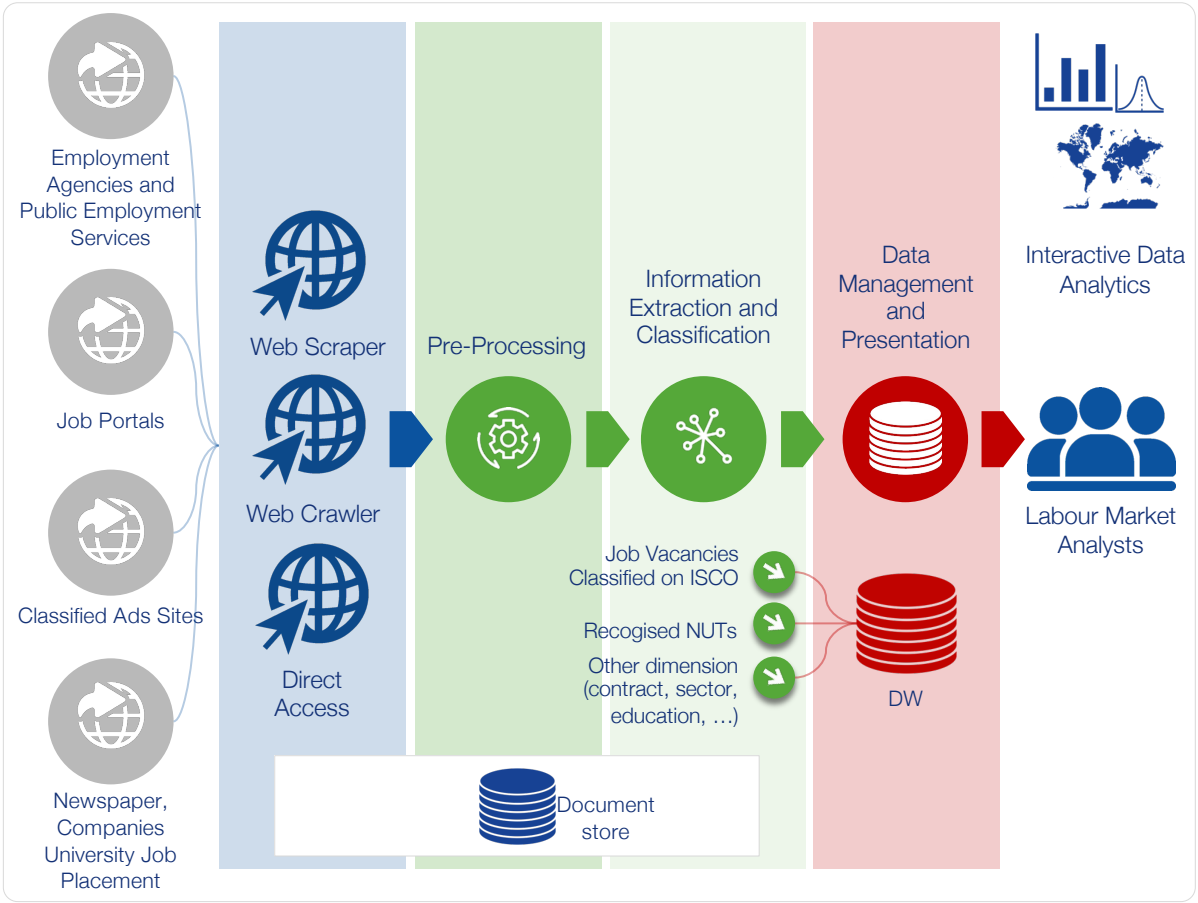
Transform Online Job Advertisments…

…in insights and analytics



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
Recogised NUTs
Other dimension (contract, sector, education, …)
Data Management and Presentation
DW
Interactive Data Analytics
Labour Market Analysts
Document store

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

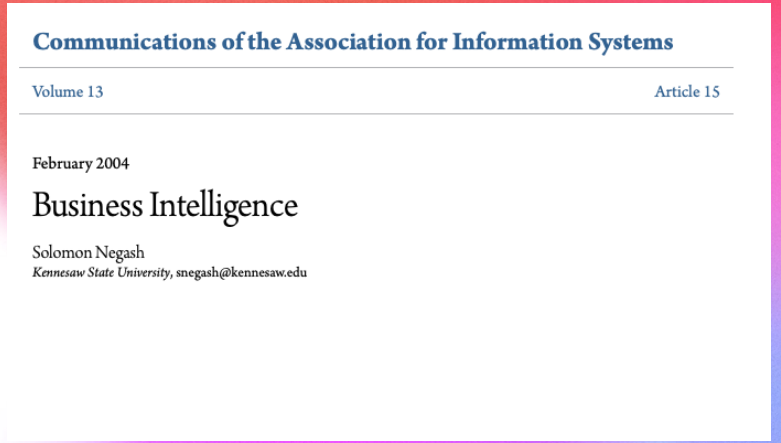
Analytics

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

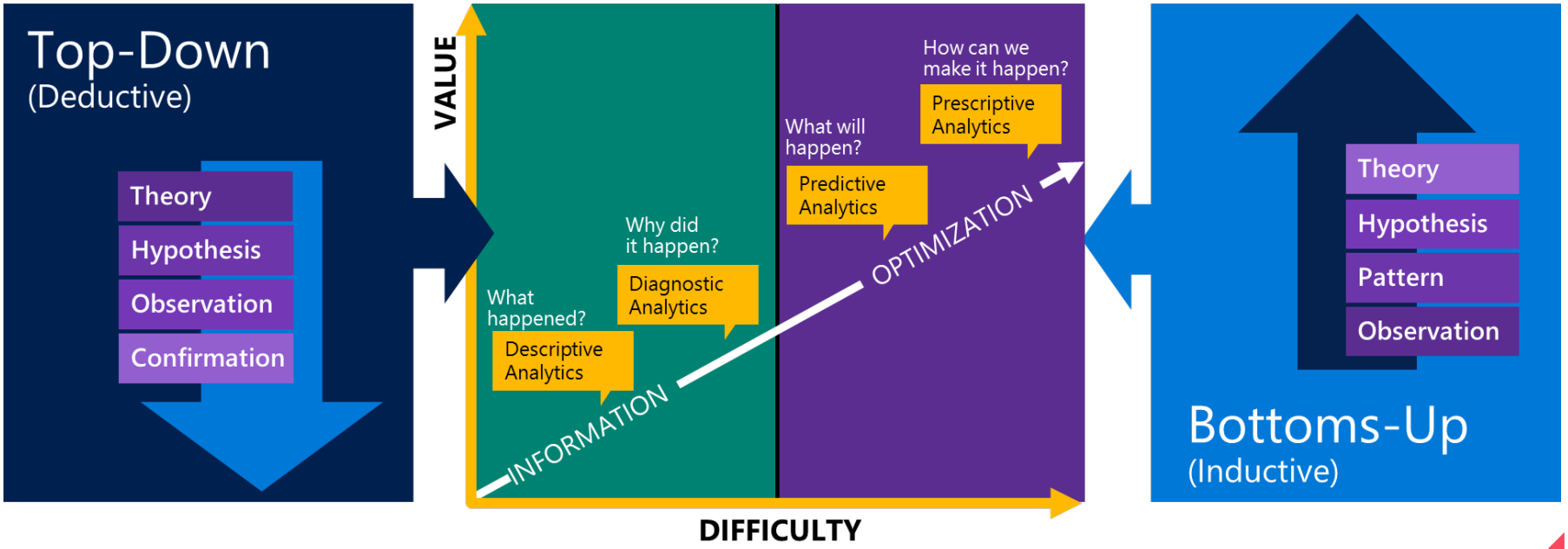
Collecting & transform

Storage

Provisioning

Application

Top-Down approach



Credits: James Serra @ Microsoft

What was the question?

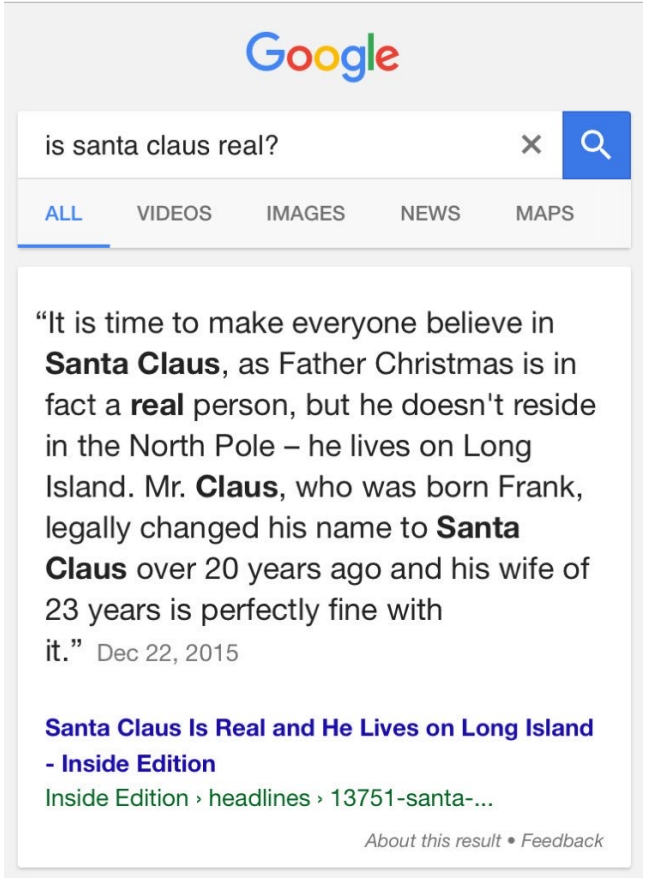
What is America's
favourite type of cake?

What was the question?

What is America's
favourite type of cake?

Favourite vs
Retail sales data

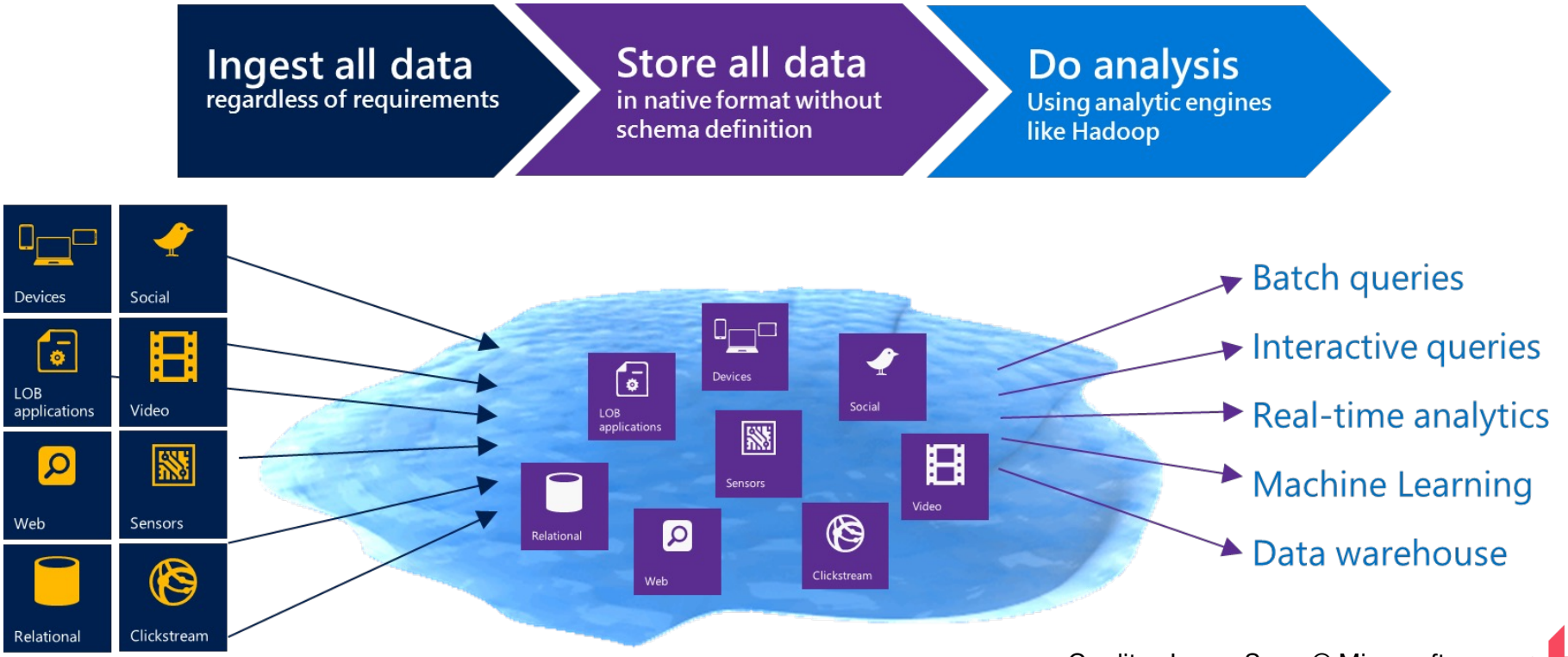
New questions



Data Warehouses meet Data Lakes

New sources

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

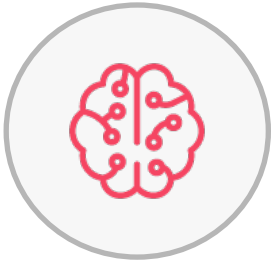
Data lake

- 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

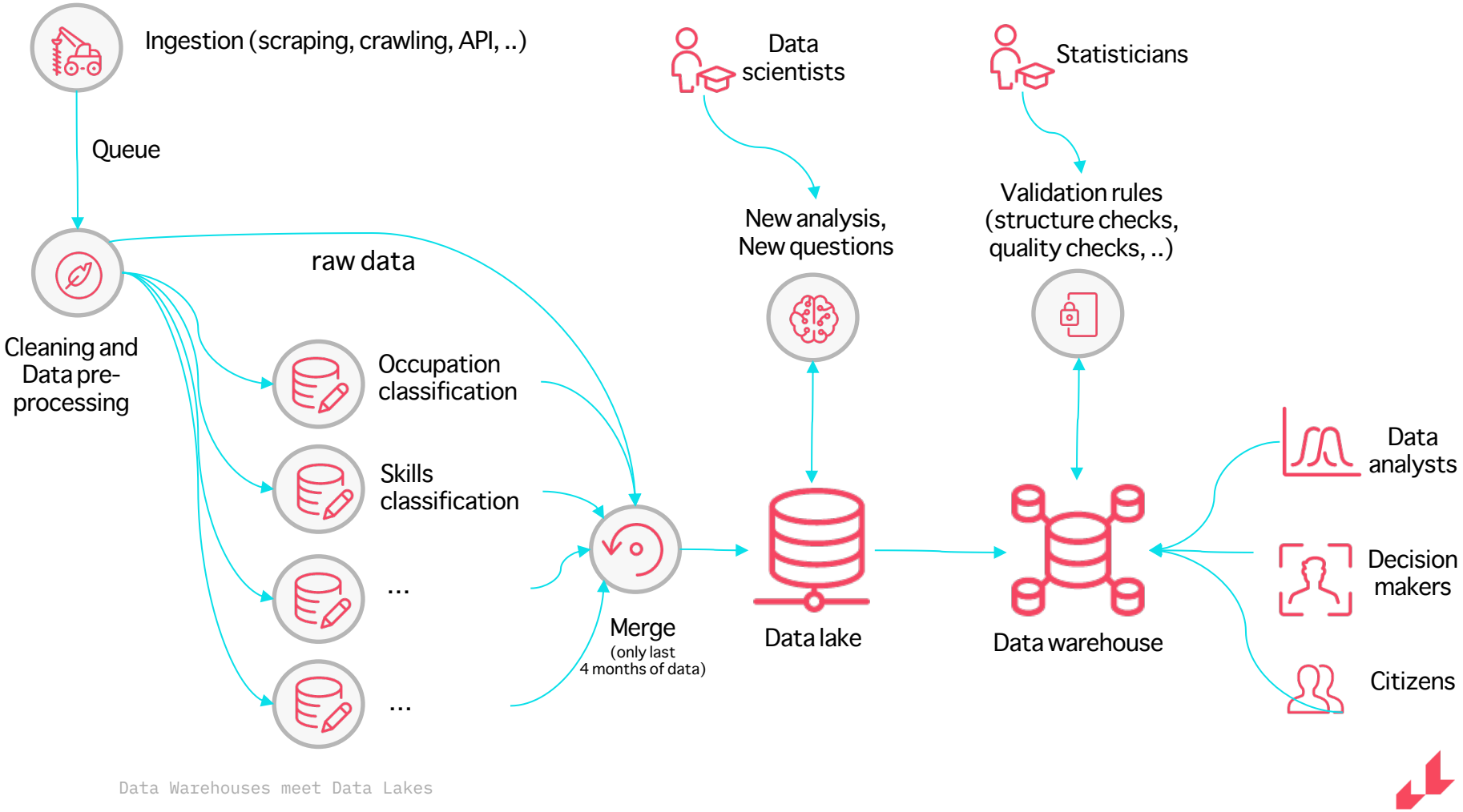
- Aggregated Subsets
- On-Demand Views
- Curated By Experts
- Structured - Tables, Views, Reports. Limited Context
- Data Quality Is Known And Tracked

Data Lake

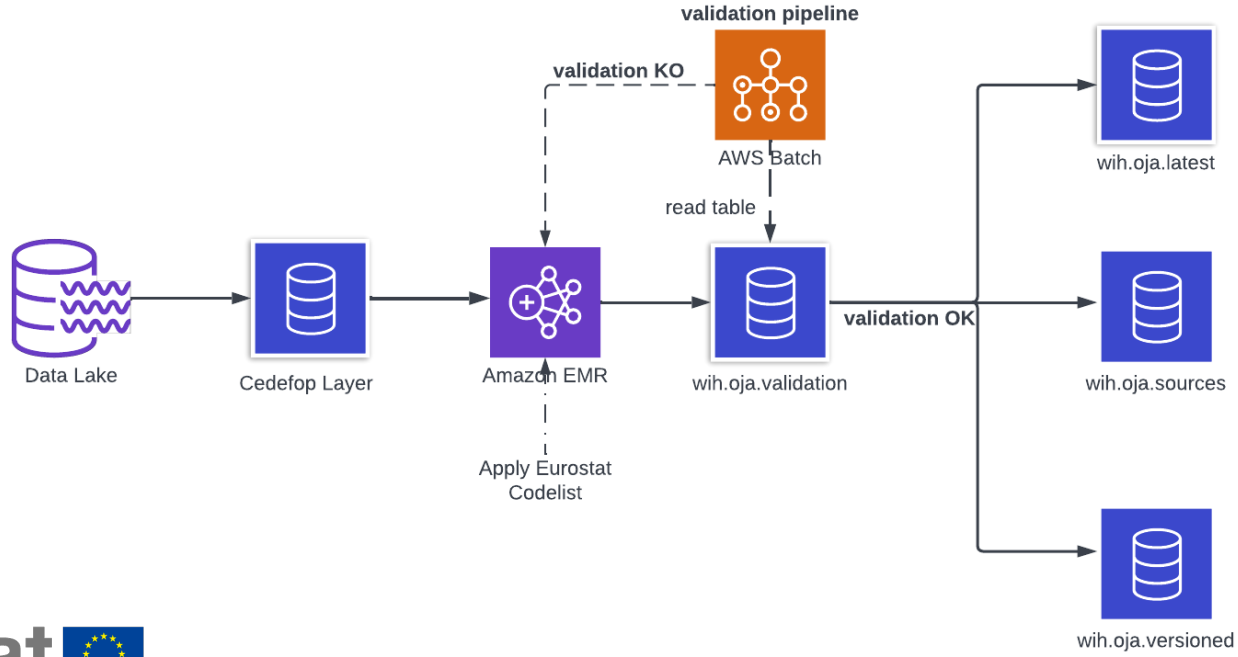
- Store Everything As-is
- Let Business Decide What They Need
- Support Rapid Change
- Provide Data Lineage and History Tracking and Visualization
- Unstructured - Key-Word Search
- Data Is Available In Various States from Raw to Fully Conformed
- Quality Metrics Often Not Available

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

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

Concepts
An Example: Columnar vs Row-Based

Row based

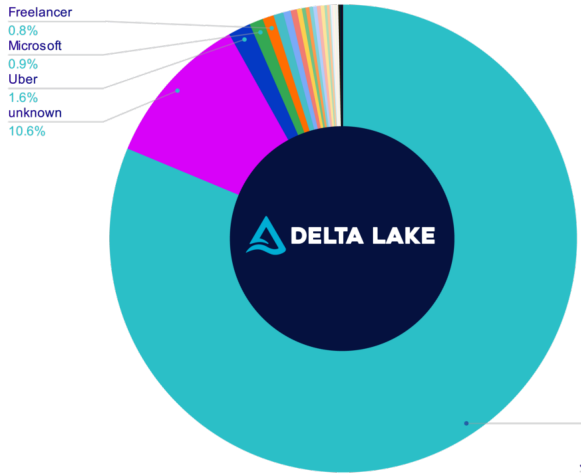
	name	color	city	age
Row 1	Tom	red	Chicago	32
Row 2	Sally	blue	Paris	87
Row 3	Mike	green	London	20
Row 4	Mary	yellow	Fresno	55

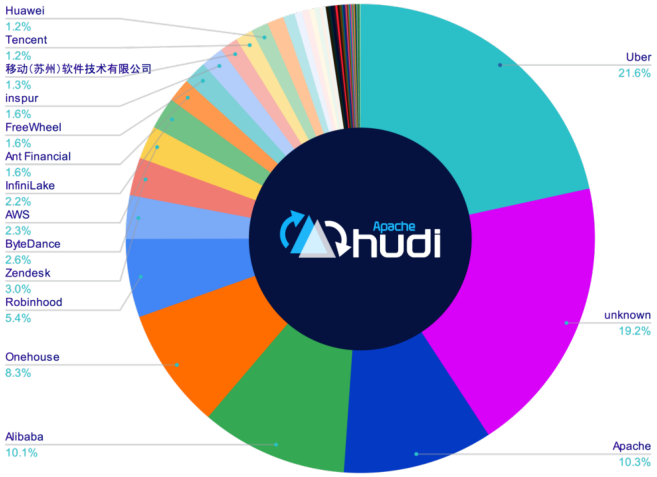
Reads Row #1

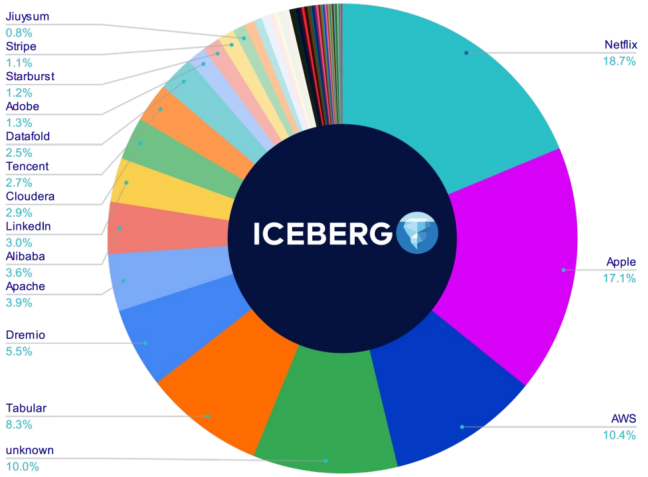
Columnar

	Row 1	Row 2	Row 3	Row 4
name	Tom	Sally	Mike	Mary
color	red	blue	green	yellow
city	Chicago	Paris	London	Fresno
age	32	87	20	55

Reads the "name" column





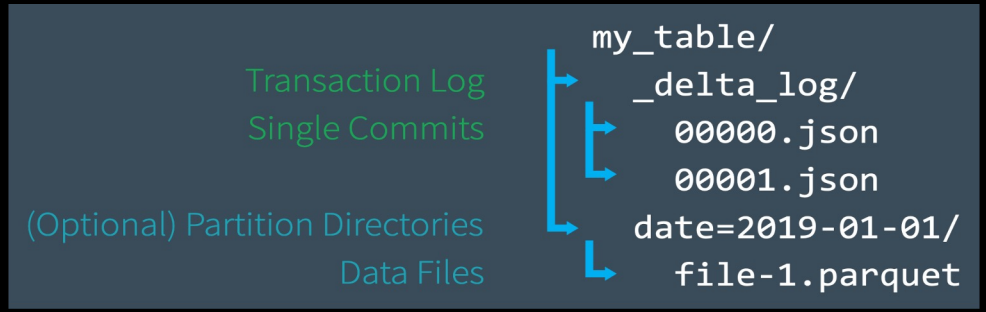


Data Warehouses meet Data Lakes

@Source dremio.com
https://www.dremio.com/subsurface/comparison-of-data-lake-table-formats-iceberg-hudi-and-delta-lake/

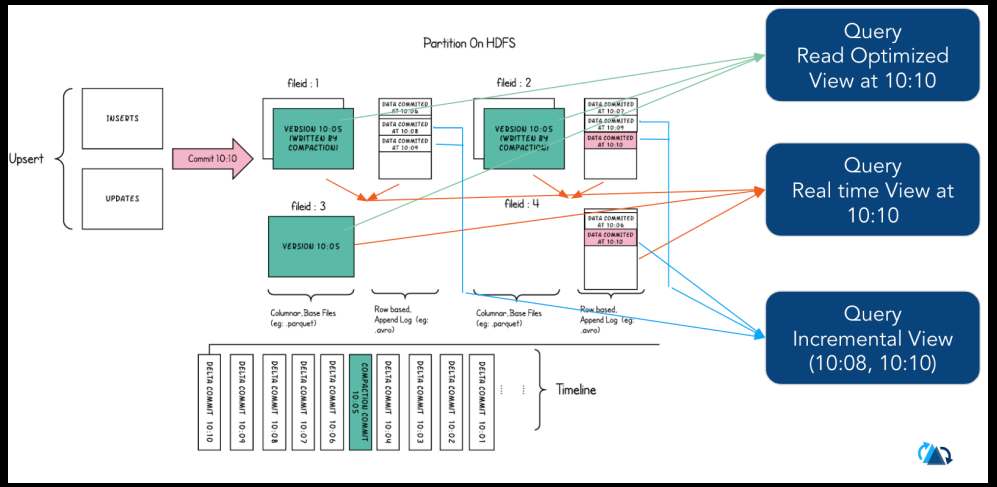
What is Delta.io?

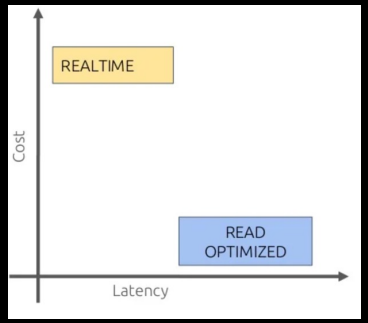
- Technology designed to be used with Apache Spark to build robust data lakes
- ACID transactions on Spark
- Scalable metadata handling
- Streaming and batch unification
- Schema enforcement
- Time travel
- Upserts and deletes
- Fully configurable/optimizable
- Structured streaming support



What is Apache Hudi?

- Schema evolution
- Handling types
- Handling guy
- Compatibility of Spark and Kafka
- Query Hudi with Athena
- Integrated in AWS



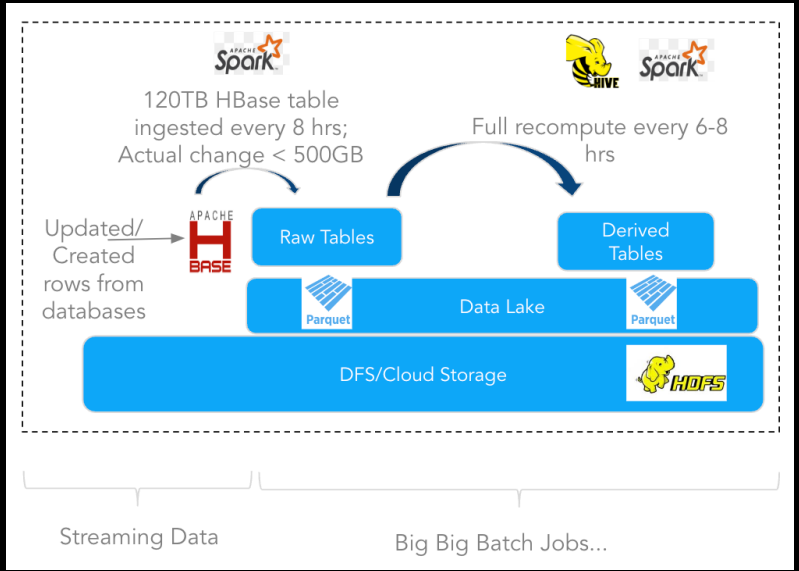


@Source:
Apache Hudi

Data Warehouses meet Data Lakes

Data Lake In-efficiency

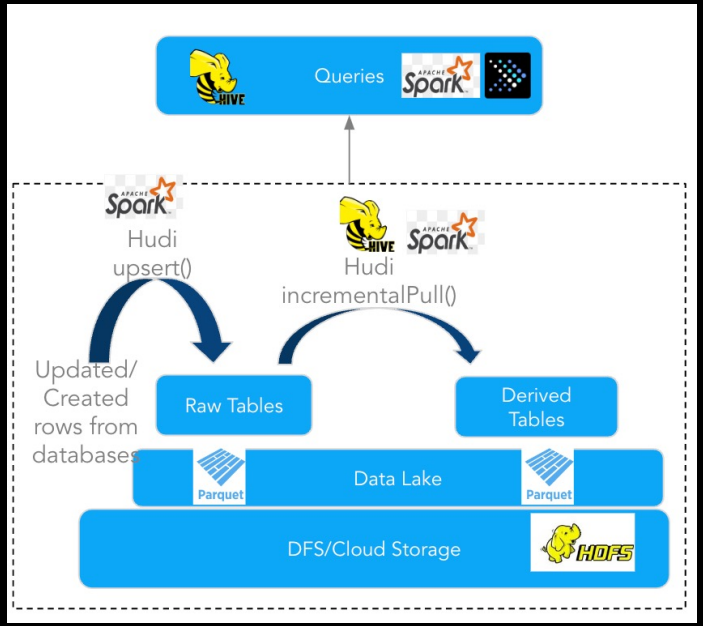
- Batch ingestion too slow
- Rewrite entire tables/partitions several times
- ETLs off raw data have no smarts to recompute
- Late arriving data is a nightmare



@Source:
Apache Hudi

Hudi Data Lake

- Design principles :
  - Let’s run mini-batch jobs, in streaming fashion. Move away from big batches.
  - Think of it as a database problem
  - Pay consideration to 10-100x more data scale & analytical workloads
  - Provide different knobs for different trade-offs



@Source:
Apache Hudi

What is Iceberg?

- Scalable format for tables
- Time travel
- No dependency from Spark
- Robust schema and partitioning changes
- Fast query planning
- Component: Catalog, Metadata layer, data layer

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

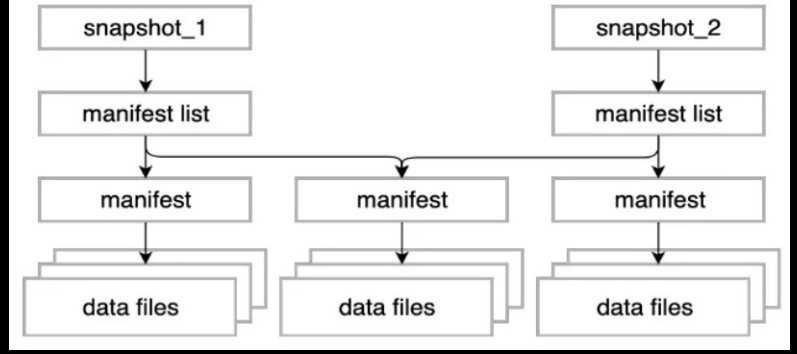


	Table Metadata	Transaction Atomicity	Isolation Levels
Delta Lake	Transaction Log + Metadata Checkpoints	Atomic Log Appends	Serializability, Strict Serializability
Hudi	Transaction Log + Metadata Table	Table-Level Lock	Snapshot Isolation
Iceberg	Hierarchical Files	Table-Level Lock	Snapshot Isolation,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

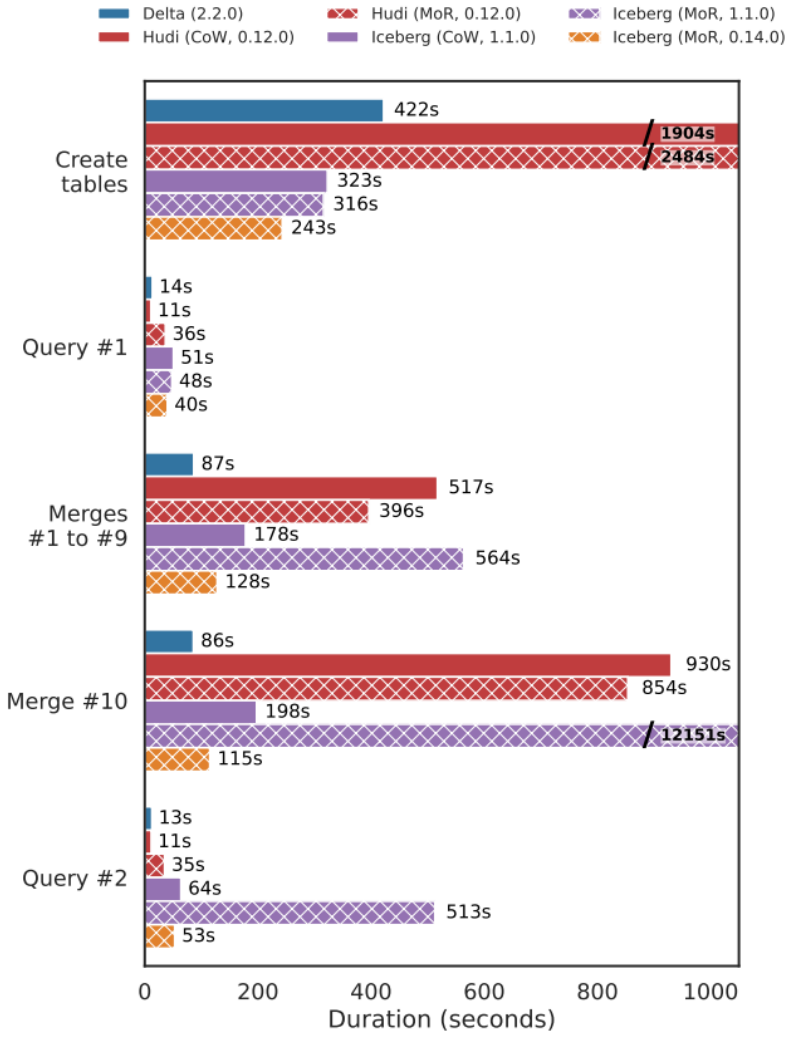


Figure 2: Performance of the 100 GB TPC-DS incremental refresh benchmark. The benchmark loads data, runs five queries (Q3, Q9, Q34, Q42, and Q59), then merges changes into the table ten times and runs these queries again. Hudi ran compaction in merge iteration 10, so we report it separately from 1–9. 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*.

Features	Delta.io – Databrics	Icerberg	Apache Hudi
Updates & deletes (ACID Transactions)	Yes	Yes	Yes
Partition Evolution	No	Yes	No
Schema Evolution	Partial	Yes	Partial
Time travel	Yes	Yes	Yes
Compactions	Manual operations, With no backup		Yes
Engine	Spark		Spark, Hive, Presto, ..
File format	Parquet	Parquest, ORC, Avro	Parquet, ORC
SQL DML	NO		NO
Write Amplification	HIGHT		LOW
Governance	Linux-Foundation Databricks	Apache Project	Apache Project
Contributions (March 2022)	16 merged PRs 43 open PRs	2241 merged PRs 275 open PRs	2880 merged PRs 160 open PRs

Thanks

Mauro.Pelucchi@gmail.com

Mauro.Pelucchi@lightcast.io