

Data Warehouses meet Data Lakes

Mauro Pelucchi
Head of Global Data Science



Description

Many organizations have migrated their data warehouses to datalake 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 2 approaches: the bottom-up of big data and the top-down of a classic data warehouse.

In this talk, I will explain the current challenges of a datalake and how we can approach a modern data architecture with the help of pyspark, hudi, delta.io or iceberg. We will see how organize data in a data lake to support real-time processing of applications and analyzes across all varieties of data sets, structured and unstructured, how provides the scale needed to support enterprise-wide digital transformation and creates one unique source of data for multiple audiences.



I am a senior data scientist and big data engineer responsible for the design
of the "**Real-Time Labour Market Information System on Skill**

Requirements" for



CEDEFOP

I currently works for



In Lightcast, I work as **Head of Global Data Science** with the goal to explore,
design and delivery innovative solutions about Labour Market Data.

In collaboration with the **University of Milano-Bicocca**, I took part in many
research projects related to the labour market intelligence systems.

I collaborate with the **University of Milano-Bicocca** as a lecturer at the
Master Business Intelligence and Big Data Analytics and with the

University of Bergamo as a lecturer in Computer Engineering.

Data Warehouses meet Data Lakes



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/europython2022>



Real-Time Labour Market Information System on Skill Requirements



The screenshot shows the Cedefop website's "Publications and reports" section. The main navigation bar includes links for Home, Themes, Publications and reports (which is highlighted), Online tools, News and events, Countries, and About Cedefop. Below this, a specific report page is displayed for the year 2019. The page title is "Real-time labour market information on skill requirements: setting up the EU system for online vacancy analysis". It features a "Global report" section and a "Country-specific report details" section. The "Country-specific report details" section includes a "Country report type: Online job vacancy market" field and a "Downloads" section containing a link to the report document. The footer of the page includes links for "Country-specific report details" and "Downloads".



Continuously evolving Labour Market

Context

Digitalization of professions
Relevance of Soft skills
Internationalisation
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



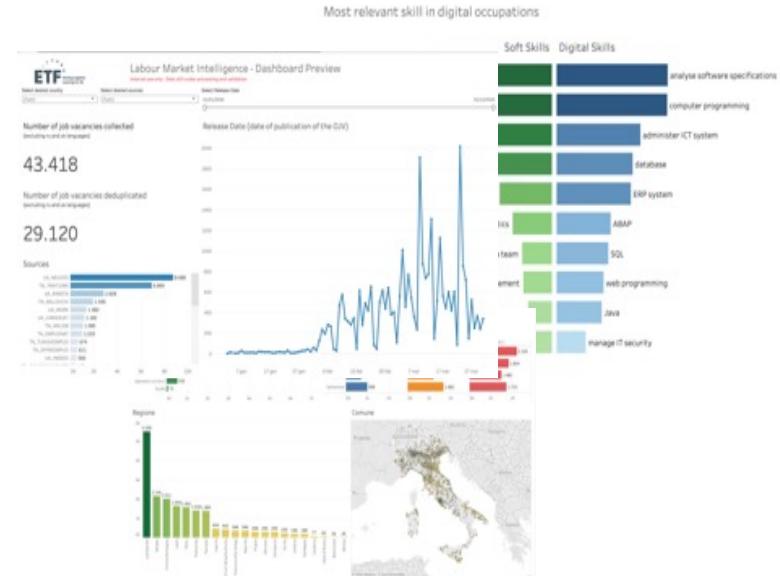
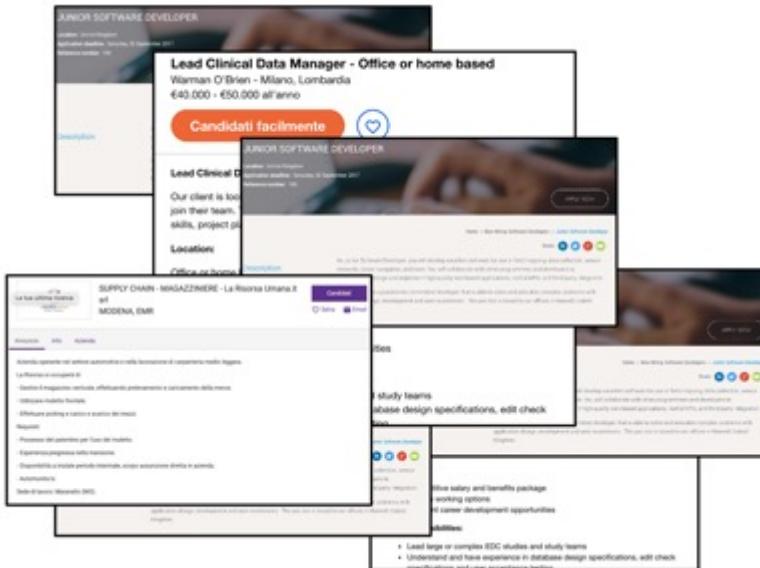
Big data

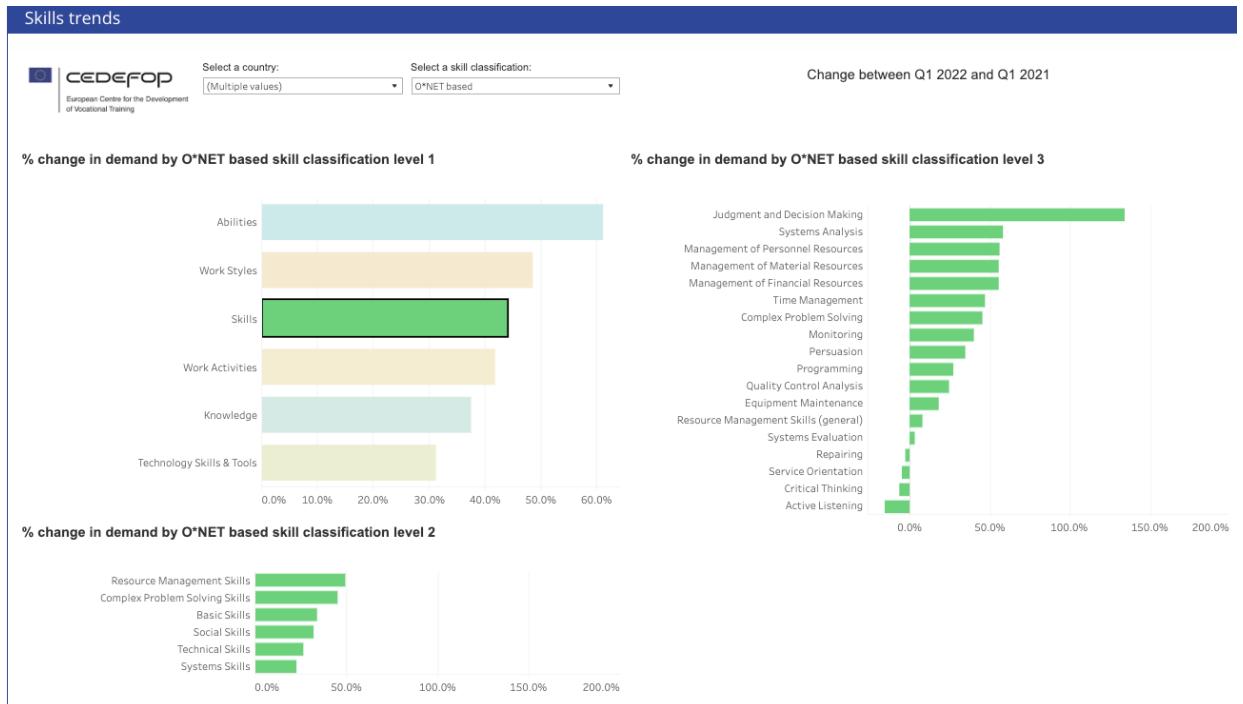
How to find a similar, complementary source of information?

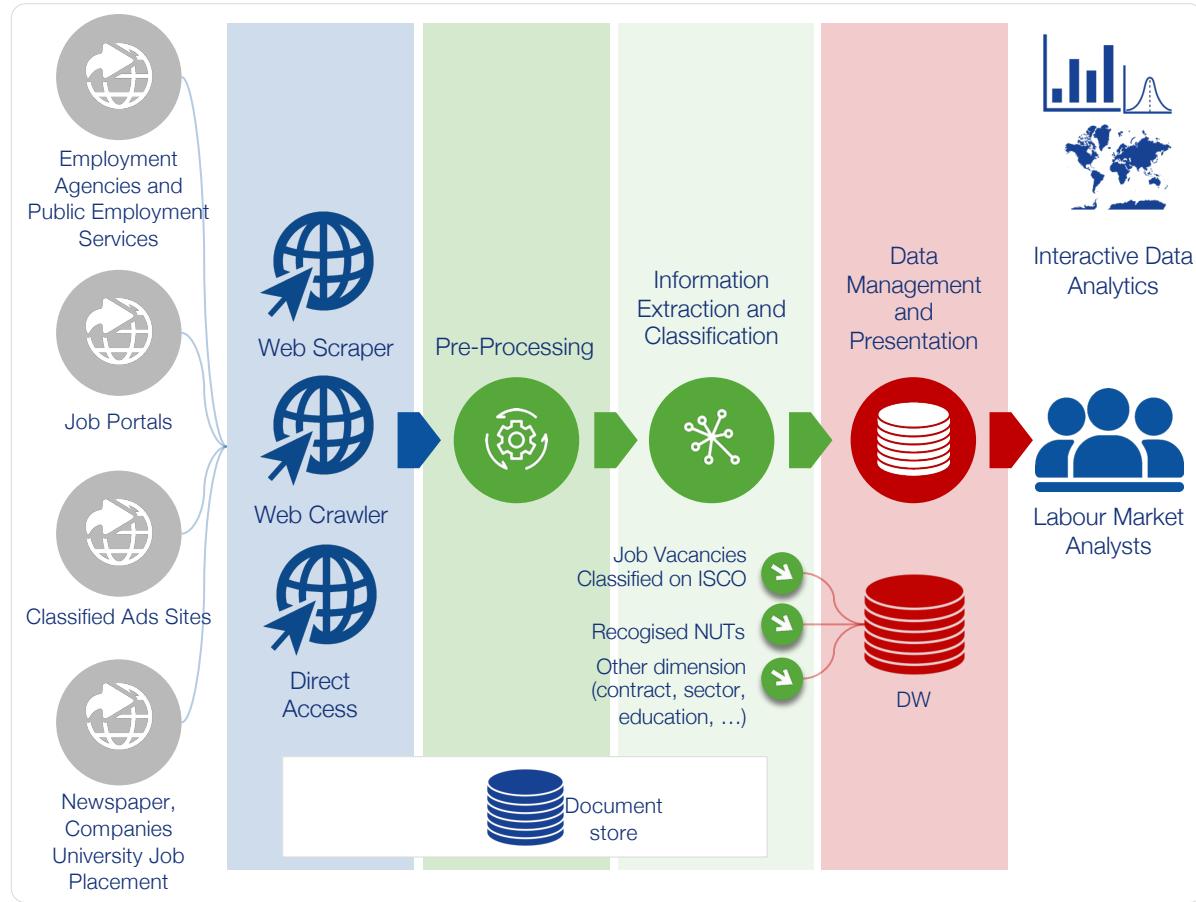
Using Web Labour Market

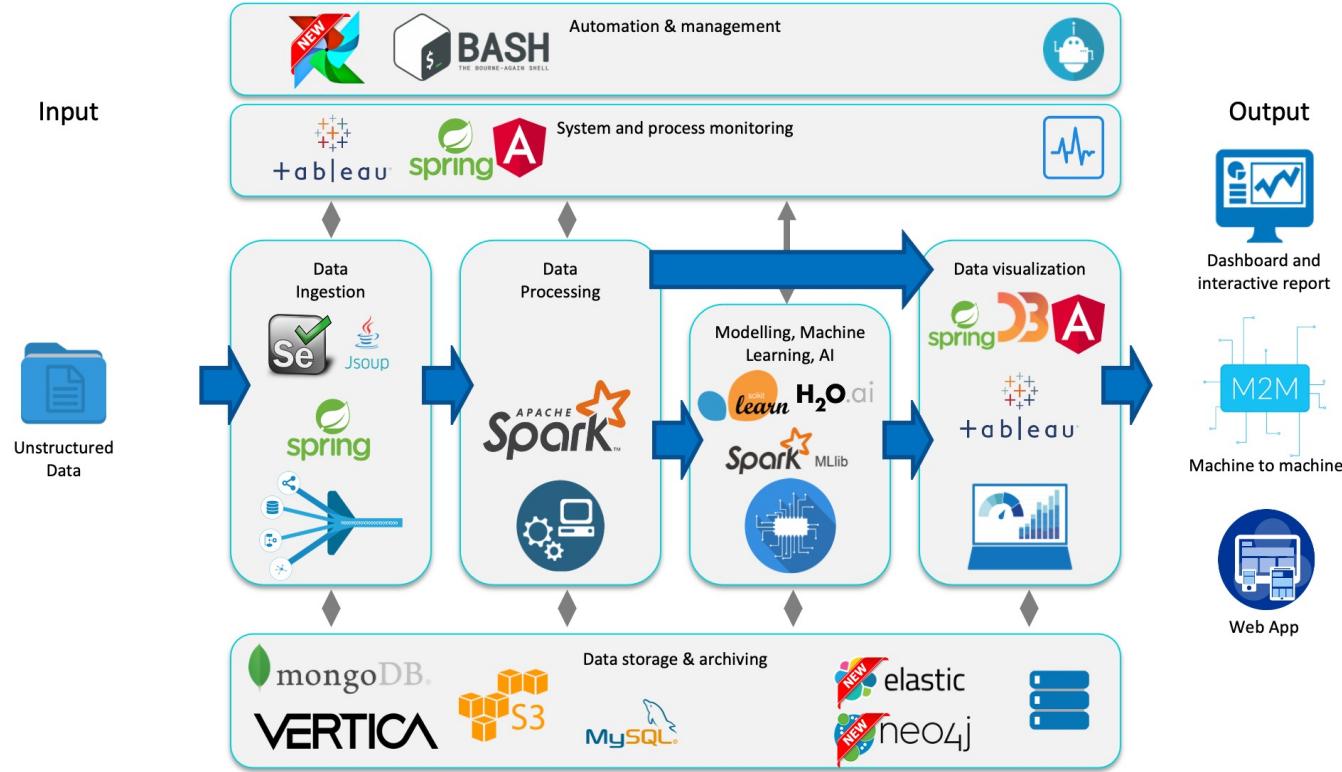


Transform Online Job Advertisements... ...in insights and analytics









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



What is America's favourite type of cake?



Apple Pie



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?*



How do we know? Business intelligence process



Retail sales data



*ETL and data
warehouse*



*Provisioning to
our marketing and
sales team*



*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.

Communications of the Association for Information Systems

Volume 13

Article 15

February 2004

Business Intelligence

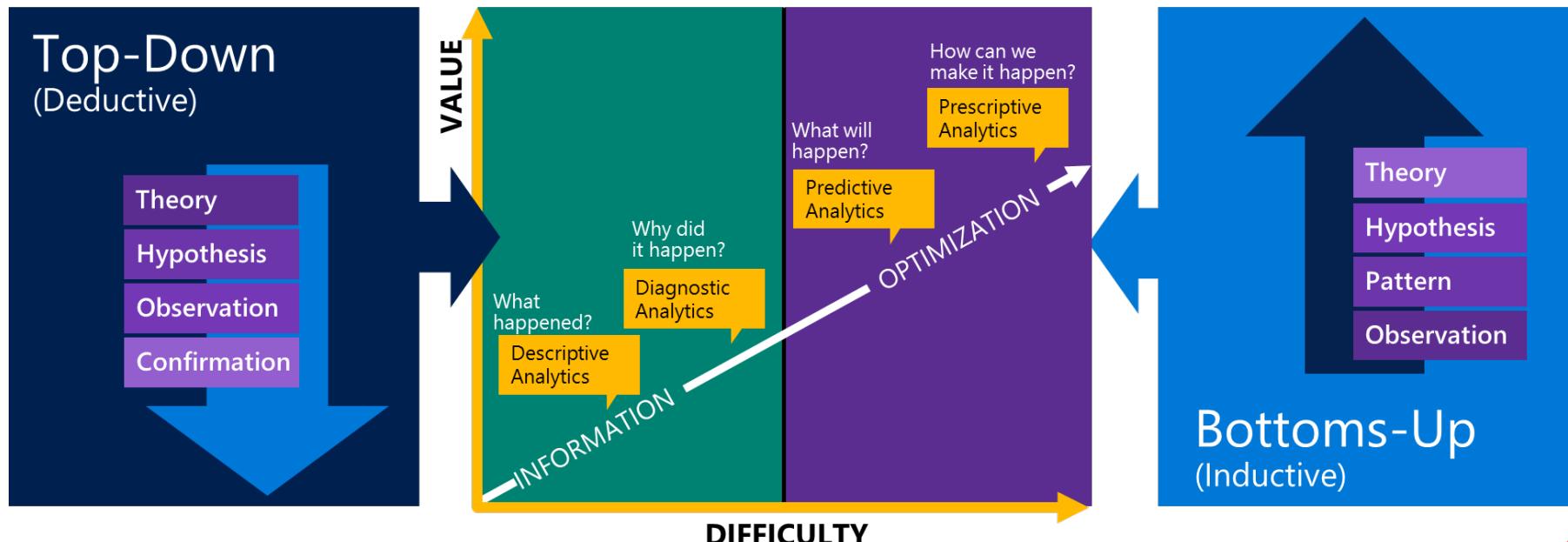
Solomon Negash

Kennesaw State University, snegash@kennesaw.edu

BI Architecture



Top-Down approach



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

Google search results for "is santa claus real?". The search bar shows the query. Below it, a snippet of a news article from Inside Edition discusses Santa Claus's residence on Long Island.

is santa claus real? X Search

ALL VIDEOS IMAGES NEWS MAPS

"It is time to make everyone believe in **Santa Claus**, as Father Christmas is in fact a **real** person, but he doesn't reside in the North Pole – he lives on Long Island. Mr. **Claus**, who was born Frank, legally changed his name to **Santa Claus** over 20 years ago and his wife of 23 years is perfectly fine with it." Dec 22, 2015

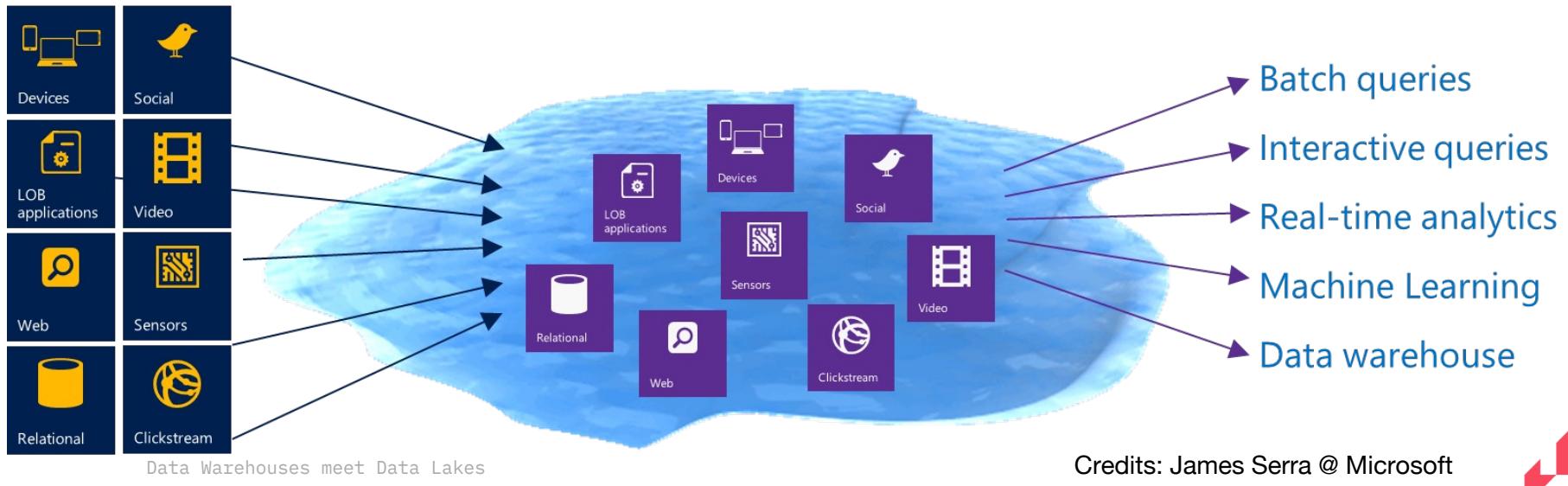
Santa Claus Is Real and He Lives on Long Island
- Inside Edition
Inside Edition › headlines › 13751-santa-...
[About this result](#) • [Feedback](#)

Data Warehouses meet Data Lakes

New sources



Bottom-up approach



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

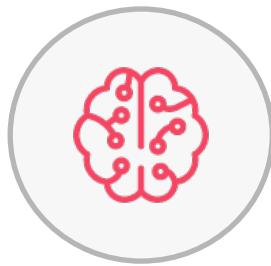
Data warehouse: collection of data (it is a database) that comes from other databases.

- 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



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

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

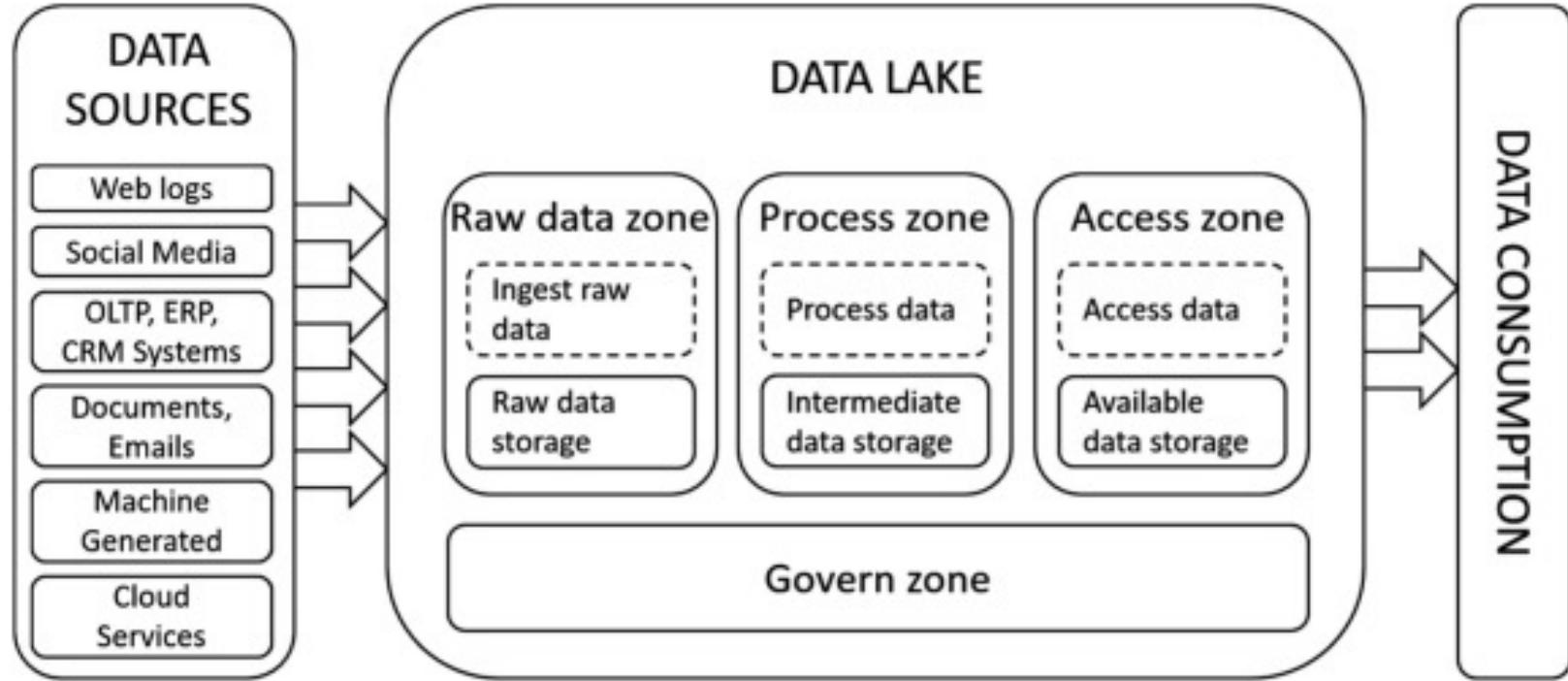
Data can be stored as is, without having to structure it first, and different types of data analysis can be performed - from control panels and visualisations to Big Data processing, real-time data analysis and machine learning to make better decisions.



Modern Day Data Lake Architecture

- Schema-on-Read
- Descriptive Data Modeling
- New Data can start flowing any time and will appear retroactively
- Flexibility
- Scalability
- Rapid Data Ingestion
- Good for Exploration and Bottom-Up Approach





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

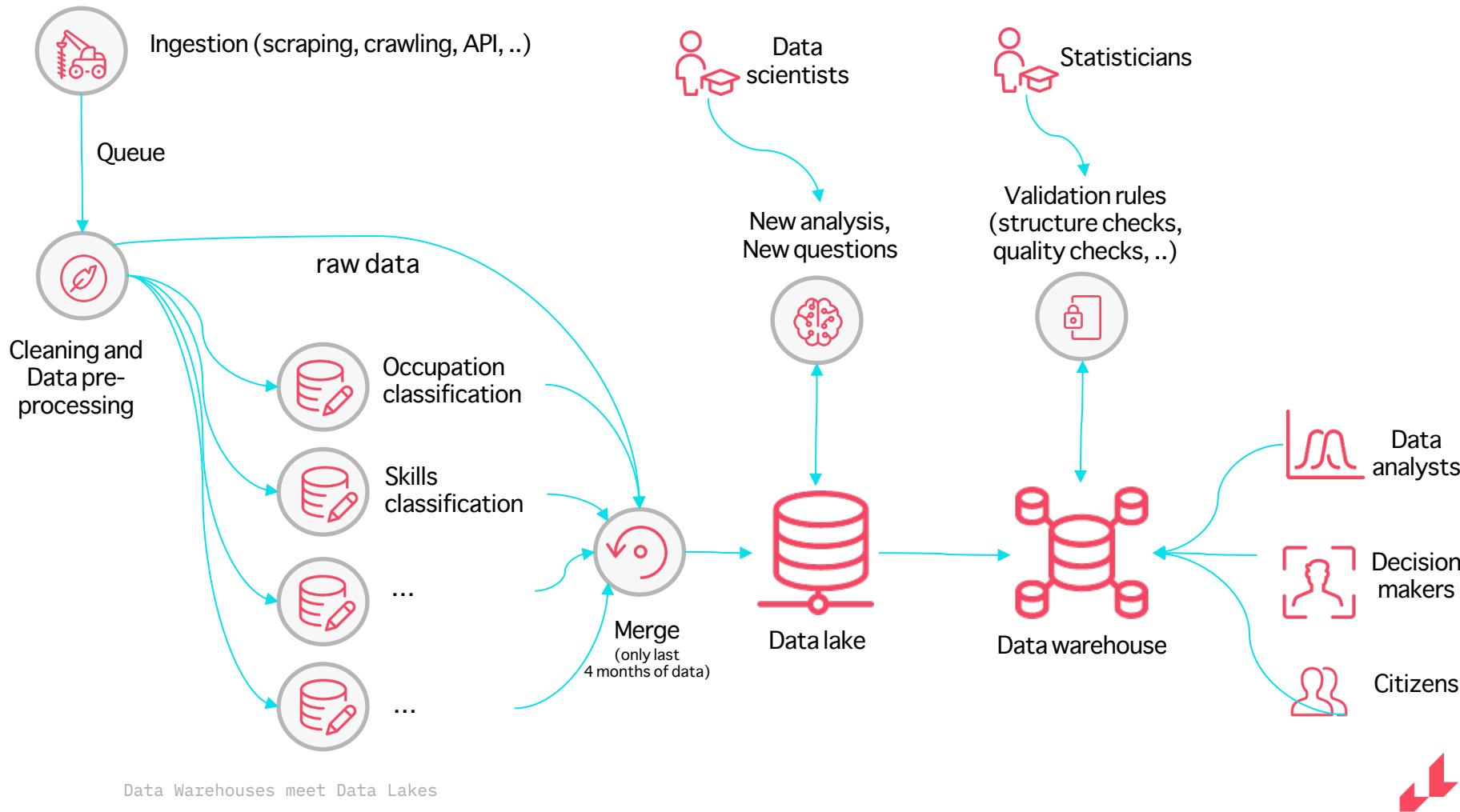
Data Lake

- 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

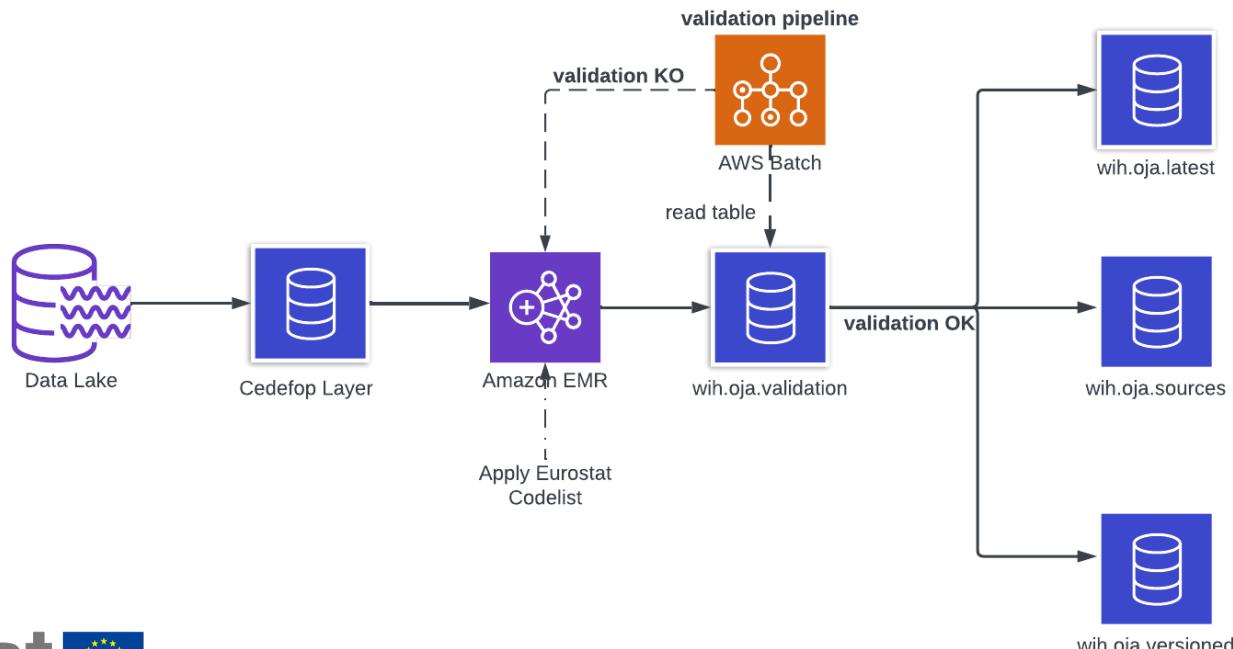


Data Lake challenge

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



Data Pipeline



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

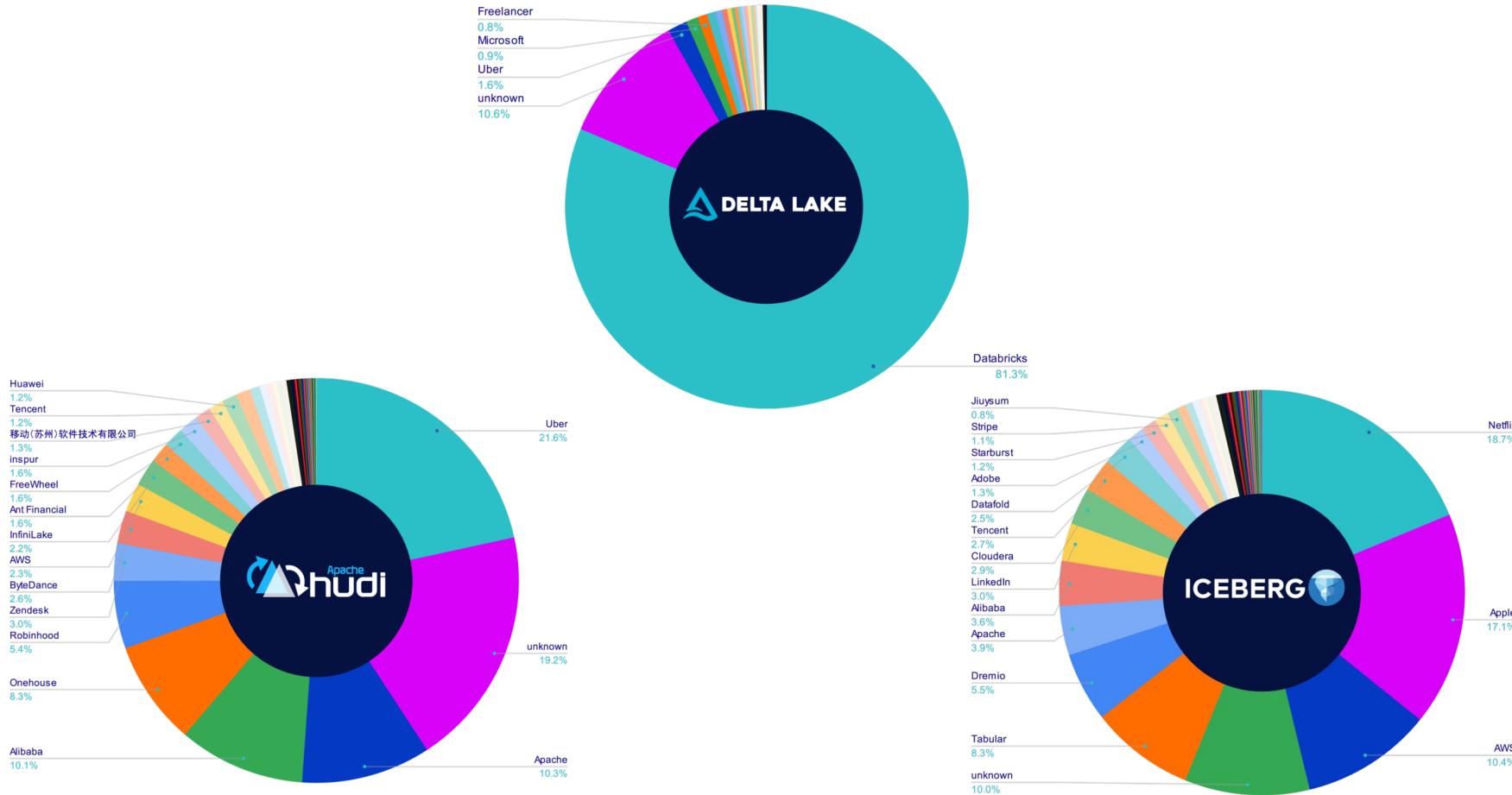
Reads Row #1

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?

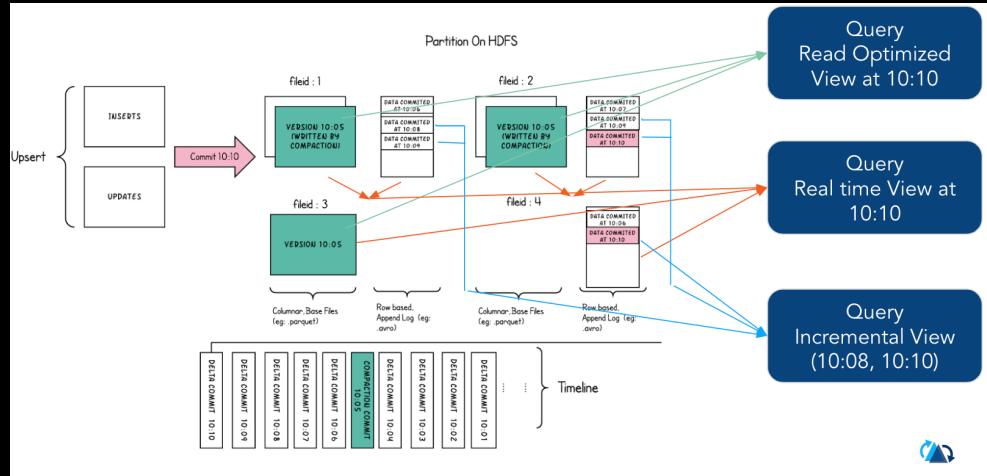
- Technology designed to be used with Apache Spark to build robust data lakes
- ACID transactions on Spark
- Scalable metadata handling
- Streaming and batch unification
- Schema enforcement
- Time travel
- Upserts and deletes
- Fully configurable/optimizable
- Structured streaming support



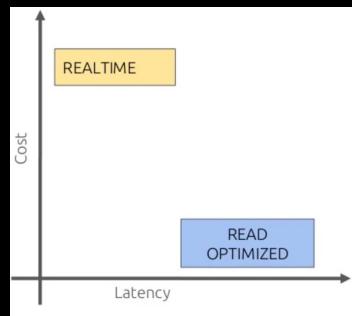
What is Apache Hudi?



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



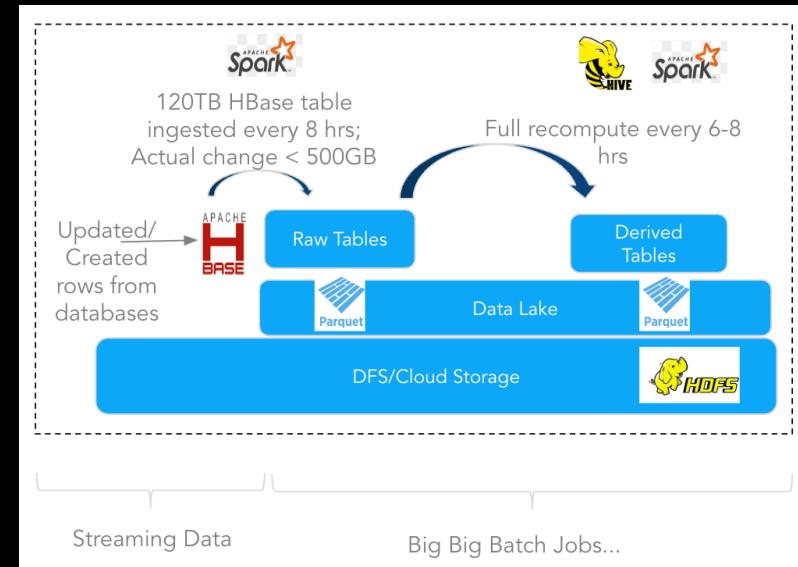
@Source:
Apache Hudi



Data Lake In-efficiency



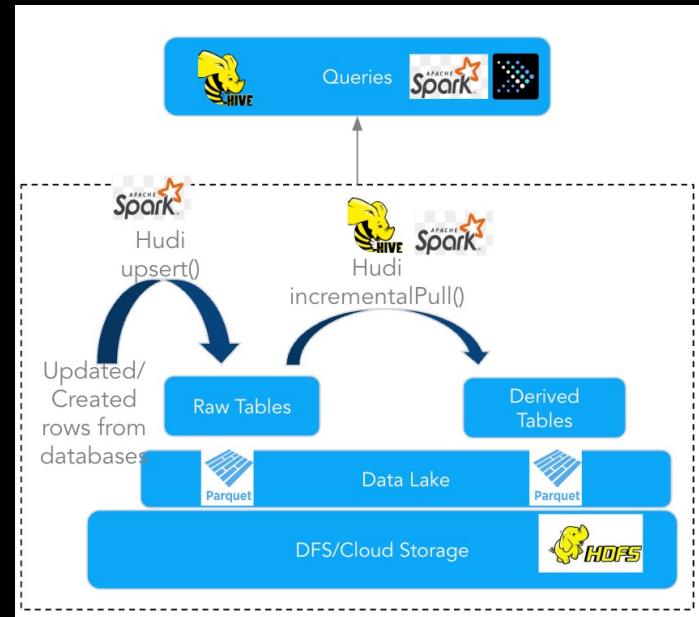
- Batch ingestion too slow
- Rewrite entire tables/partitions several times
- ETLs off raw data have no smarts to recompute
- Late arriving data is a nightmare



Hudi Data Lake



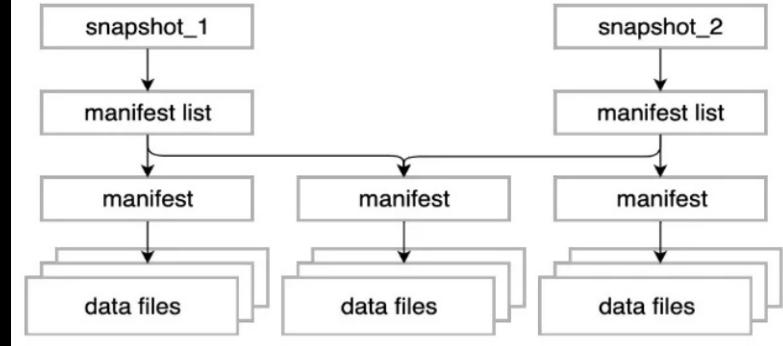
- Design principles :
 - Let's run mini-batch jobs, in streaming fashion. Move away from big batches.
 - Think of it as a database problem
 - Pay consideration to 10-100x more data scale & analytical workloads
 - Provide different knobs for different trade-offs



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")
```



Hudi vs Delta.io vs Icerberg

Features	Delta.io – Databricks	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

Thanks



Mauro.Pelucchi@gmail.com

Mauro.Pelucchi@lightcast.io

