Data Lakes and Enterprise Data Infrastructure

JIAN PEI

SIMON FRASER UNIVERSITY

Data in Enterprises

Rich

Domain knowledge accumulated

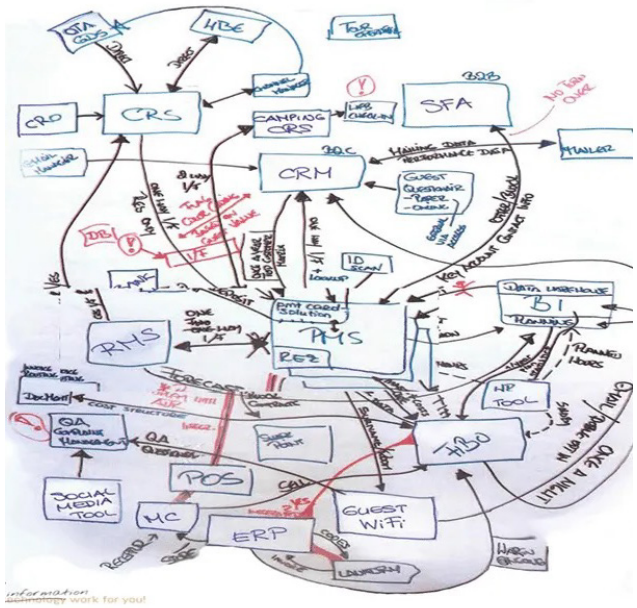
Heterogeneous

Dirty

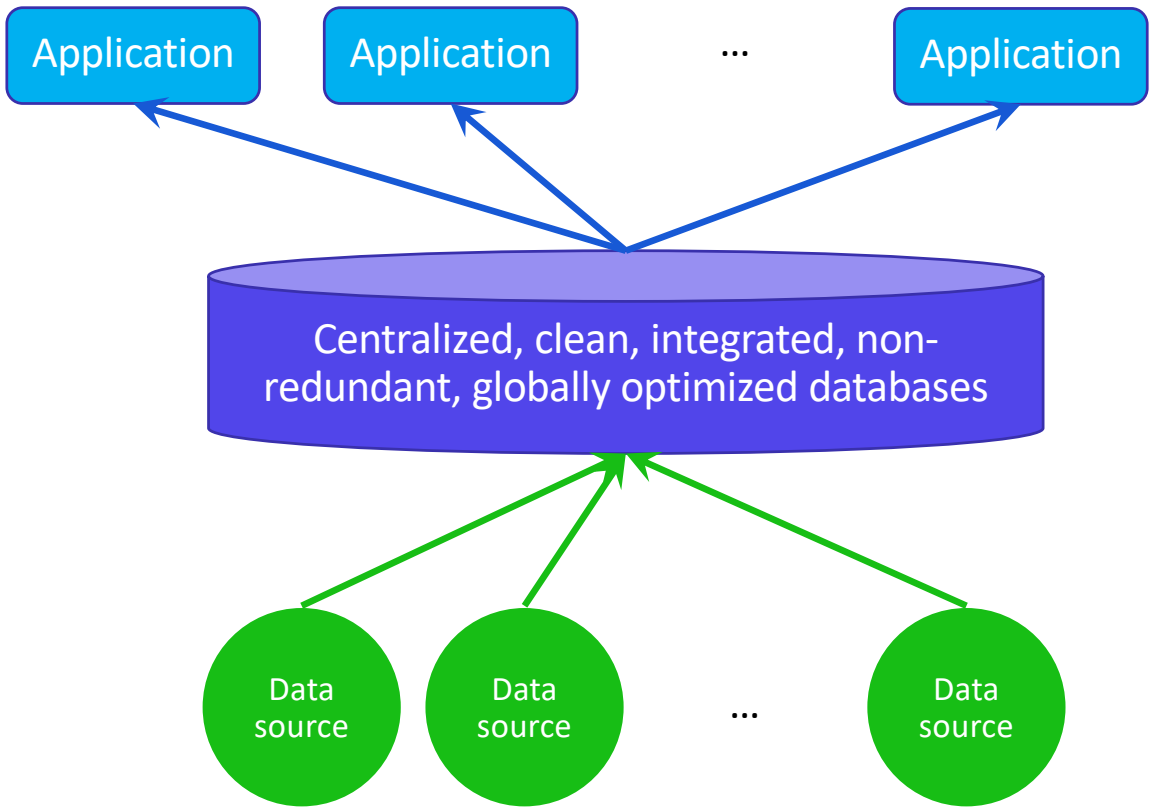
Isolated

Redundant

Locally self-ruling



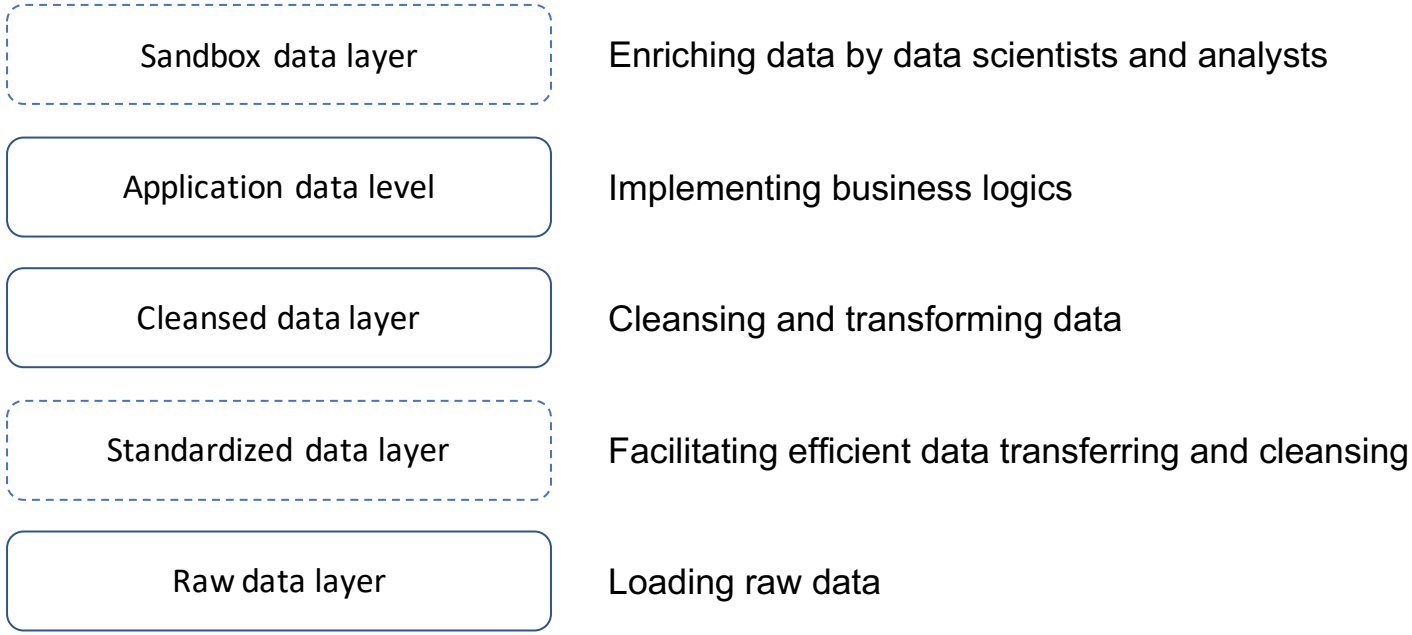
Utopia of Data Infrastructure in Enterprises



Data Lake

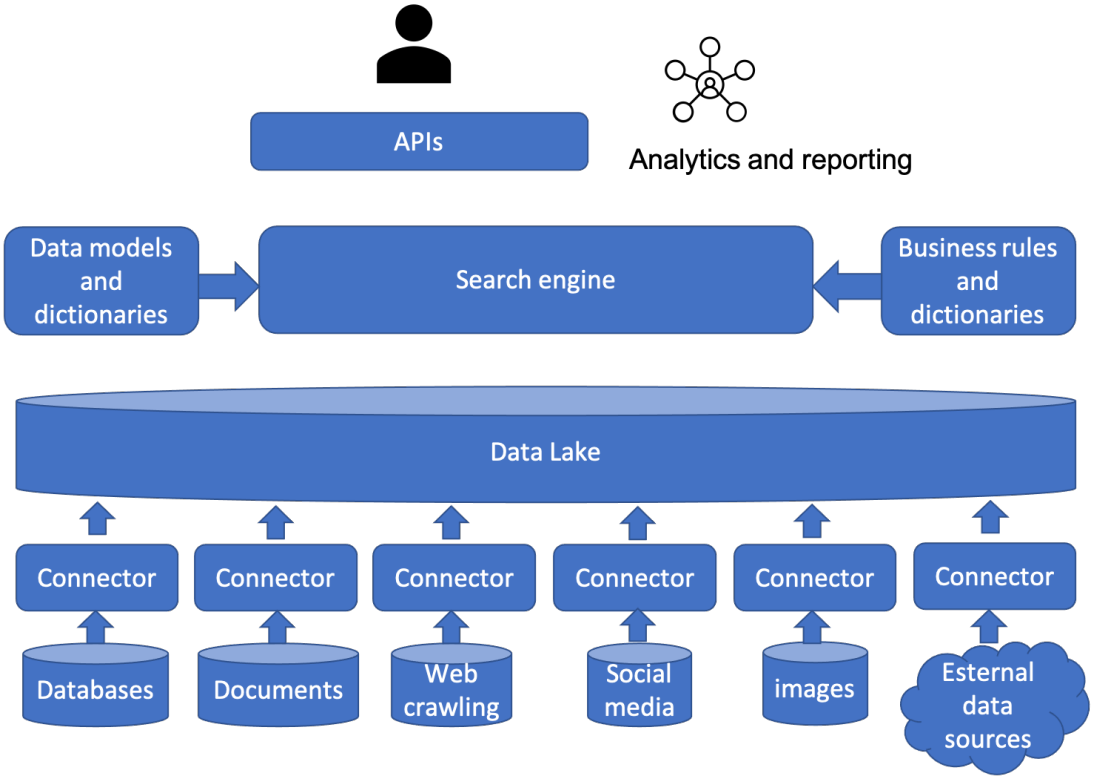
A data lake is a centralized repository storing all structured and unstructured data at any scale in an organization

- Data is stored as is, without having to first structure the data, and run different types of analytics—from dashboards and visualizations to big data processing, real-time analytics, and machine learning to guide better decision

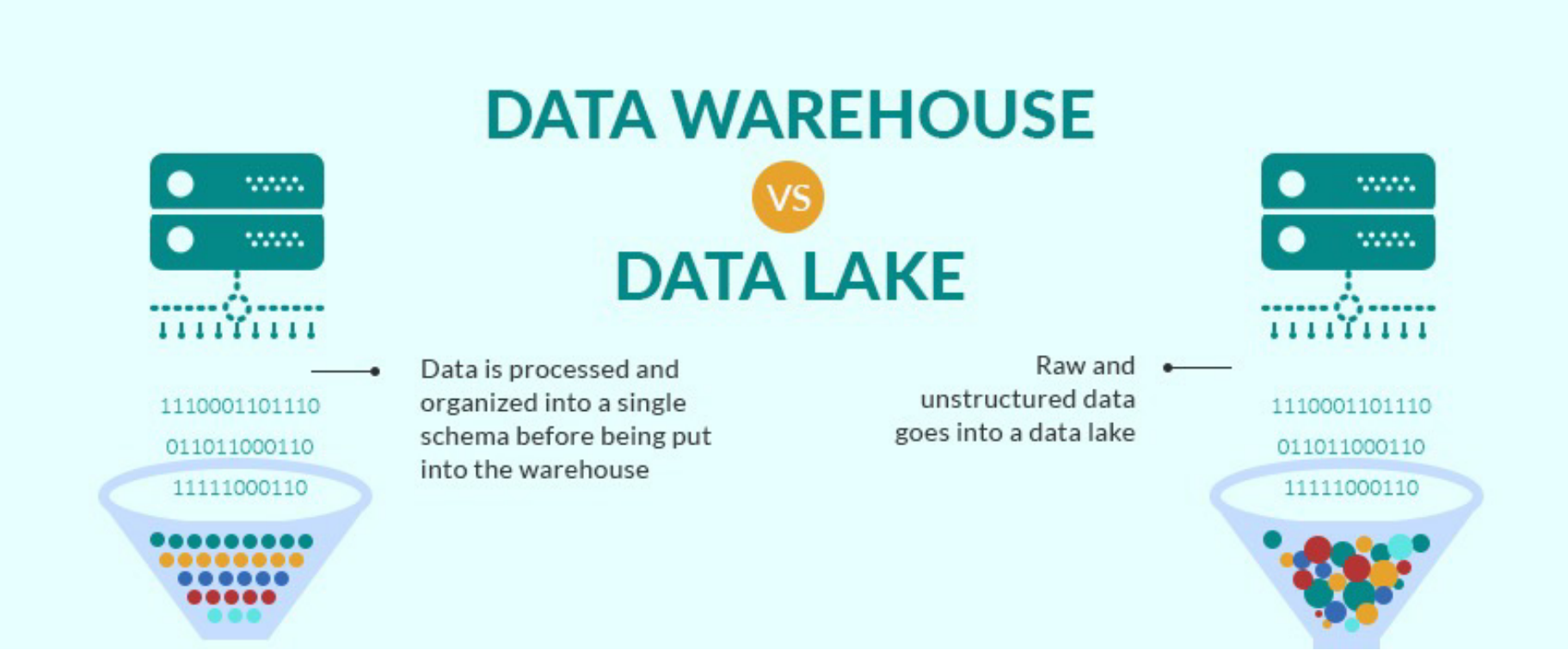


Layers of Storage

Conceptual Architecture



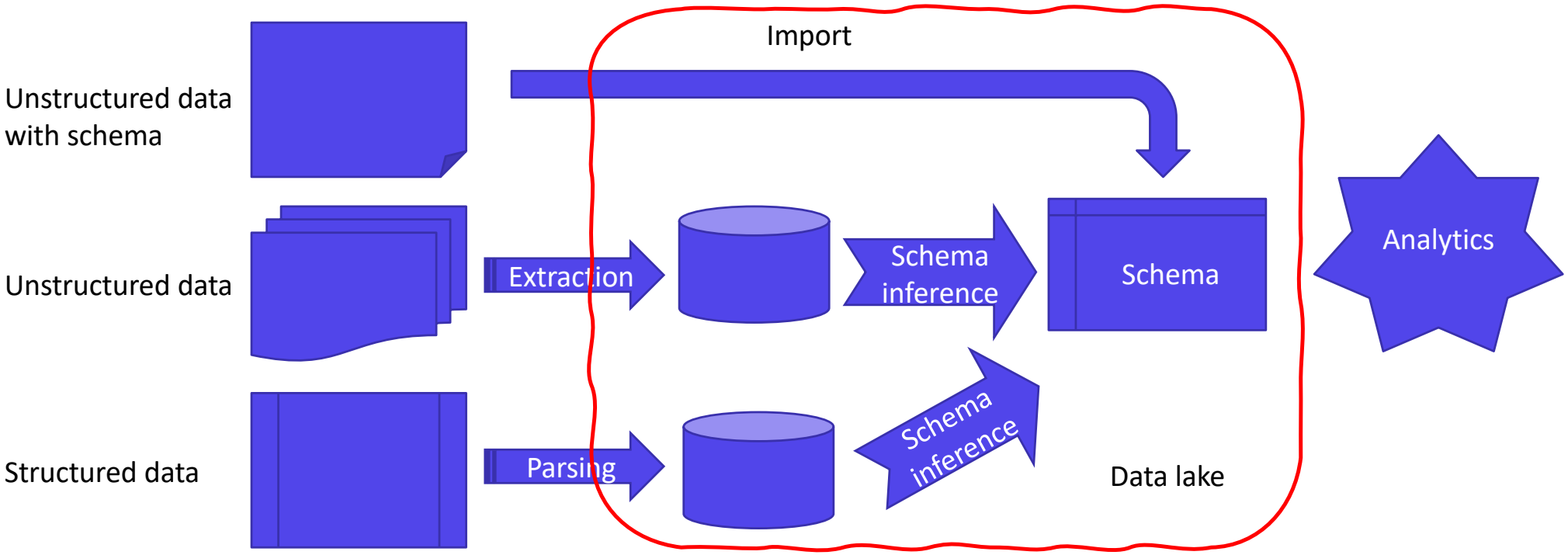
Data Warehouse versus Data Lake



Characteristics	Data Warehouse	Data Lake
Data	Relational from transactional systems, operational databases, and line of business applications	Non-relational and relational from IoT devices, web sites, mobile apps, social media, and corporate applications
Schema	Designed prior to the DW implementation (schema-on-write)	Written at the time of analysis (schema-on-read)
Price/Performance	Fastest query results using higher cost storage	Query results getting faster using low-cost storage
Data Quality	Highly curated data that serves as the central version of the truth	Any data that may or may not be curated (ie. raw data)
Users	Business analysts	Data scientists, Data developers, and Business analysts (using curated data)
Analytics	Batch reporting, BI and visualizations	Machine Learning, Predictive analytics, data discovery and profiling

Data Warehouse versus Data Lake

(Enterprise) Data Organization



Data Lake Challenges

Many data sets, hundreds of thousands and more

Complicated queries: keyword queries, finding joinable data sets, finding features, …

Core task in building data lake: meta data management

What are “VECTOR” and “COORDINATE”?

Homicides	UOM	UOM_ID	SCALAR_FACTOR	SCALAR_ID	VECTOR	COORDINATE	VALUE	STATUS	SYMBOL	TERMINATED	DECIMALS
Number of homicide victims	Number	223	units	0	v1489206	1.6.1	283				0
Percentage of homicides	Percentage	242	units	0	v1489207	1.6.2	48.05				2
Number of homicide victims	Number	223	units	0	v1489196	1.1.1	76				0
Pe						1.1.2	28.36				2
Nu						1.2.1	180				0
Pe						1.2.2	67.16				2
Nu						1.3.1		..			0
Pe						1.3.2		..			2
Nu						1.4.1	12				0
Pe						1.4.2	4.48				2

Number and percentage of homicide victims, by type of firearm used to commit the homicide

Number and percentage of homicide victims, by type of firearm used to commit the homicide (total firearms; handgun; rifle or shotgun; fully automatic firearm; sawed-off rifle or shotgun; firearm-like weapons; other firearms, type unknown), Canada, 1974 to 2018.

Publisher - Current Organization Name: Statistics Canada

Licence: Open Government Licence - Canada

Not helpful!

https://open.canada.ca/data/en/dataset/be073ee2-a302-4d32-af20-a48f5fbe2e63

Data Cleaning in Data Lakes

Data cleaning is the No. 1 most cited task in data lake, and >85% considered it either major or critical to the business.

tpep_dropoff_dat...	passenger_count	trip_...	Rate...	store...	PULo...	DOLo...	pay...	fare_amount
2017 Jan 09 11:25:45 AM	1	3.3	1	N	263	161	1	12.5
2017 Jan 09 11:36:01 AM	1	0.9	1	N	186	234	1	5
2017 Jan 09 11:42:05 AM	1	1.1	1	N	164	161	1	5.5
2017 Jan 09 11:57:36 AM	1	1.			36	75	1	6
2017 Jun 20 10:39:16 PM	1	0.			41	141	2	630,461.82
2017 Jan 19 09:29:44 AM	3				39	264	2	625,900.8
2017 Jan 01 02:57:09 AM	1	0	1	N	232	243	3	538,579.2
2017 Apr 01 07:45:43 P...	1	0	1	N	90	264	2	538,481.03
2017 Oct 10 04:33:07 P...	1	1,178.6	1	N	170	170	2	404,093.97

Are these values correct?

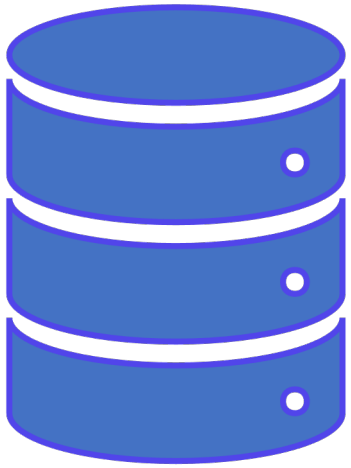
https://data.cityofnewyork.us/Transportation/**2017-Yellow-Taxi-Trip-Data**/biws-g3hs

Evolving Data in Data Lakes

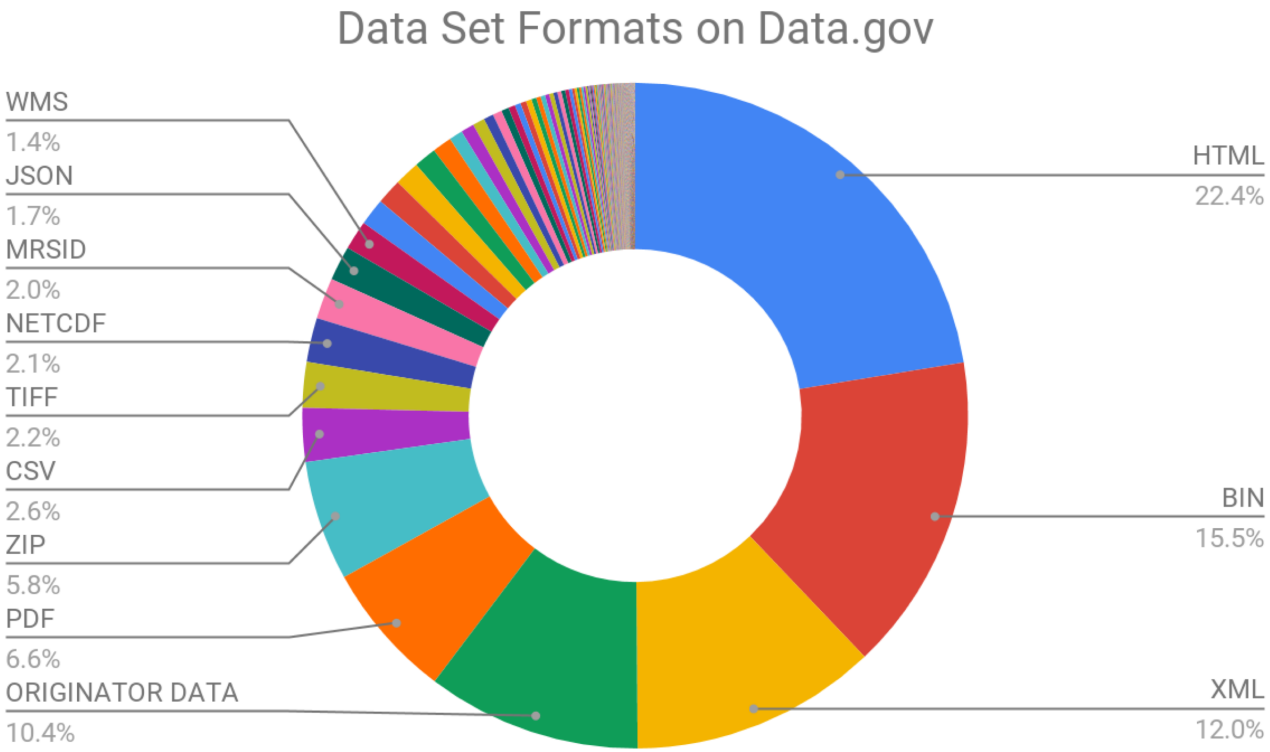
Many duplicates as data sets are often being copied for new project

Data sets are constantly being updated, having their schema altered, being derived into new ones, and disappearing/reappearing

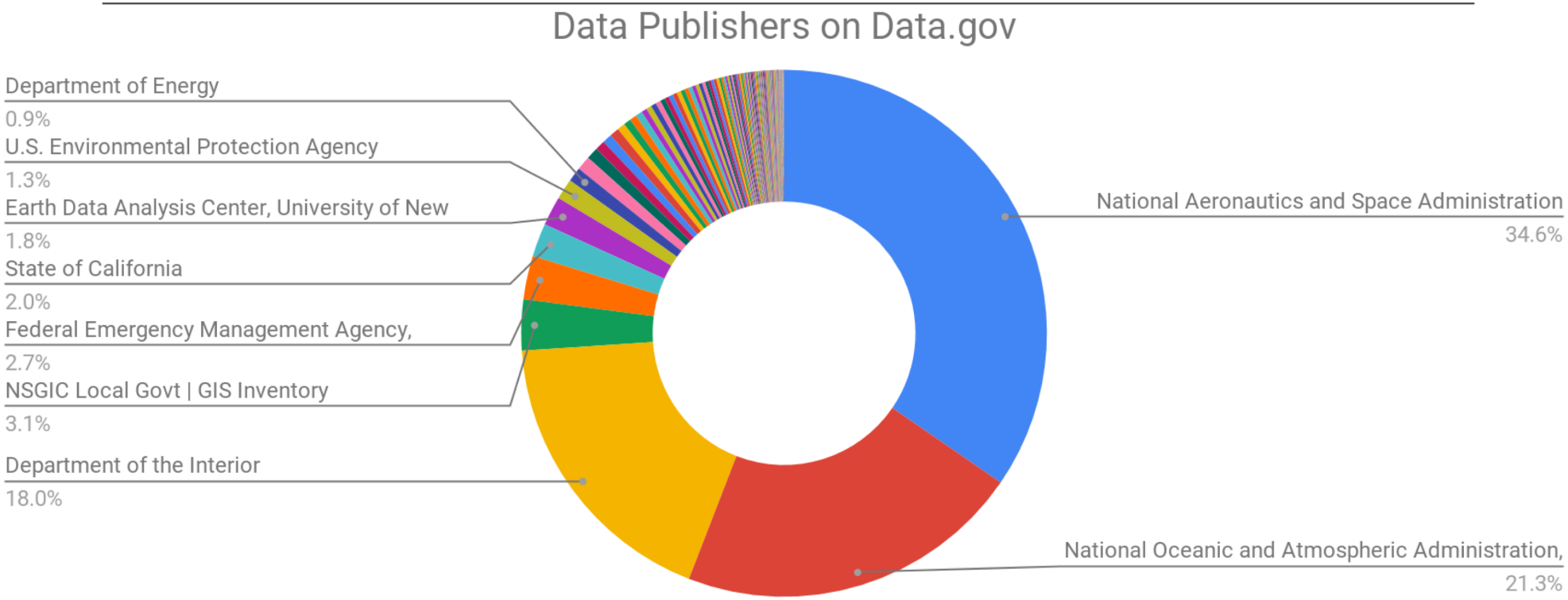
Data set versioning is to maintain all versions of datasets for storage cost-saving, collaboration, auditing, and experimental reproducibility



Diversity in Data Lakes



Diversity in Data Lakes



Diversity in Data Lakes

Dataset formats in the open world can be highly heterogeneous

Ingestion and extraction is the task of bringing structured datasets into data lake

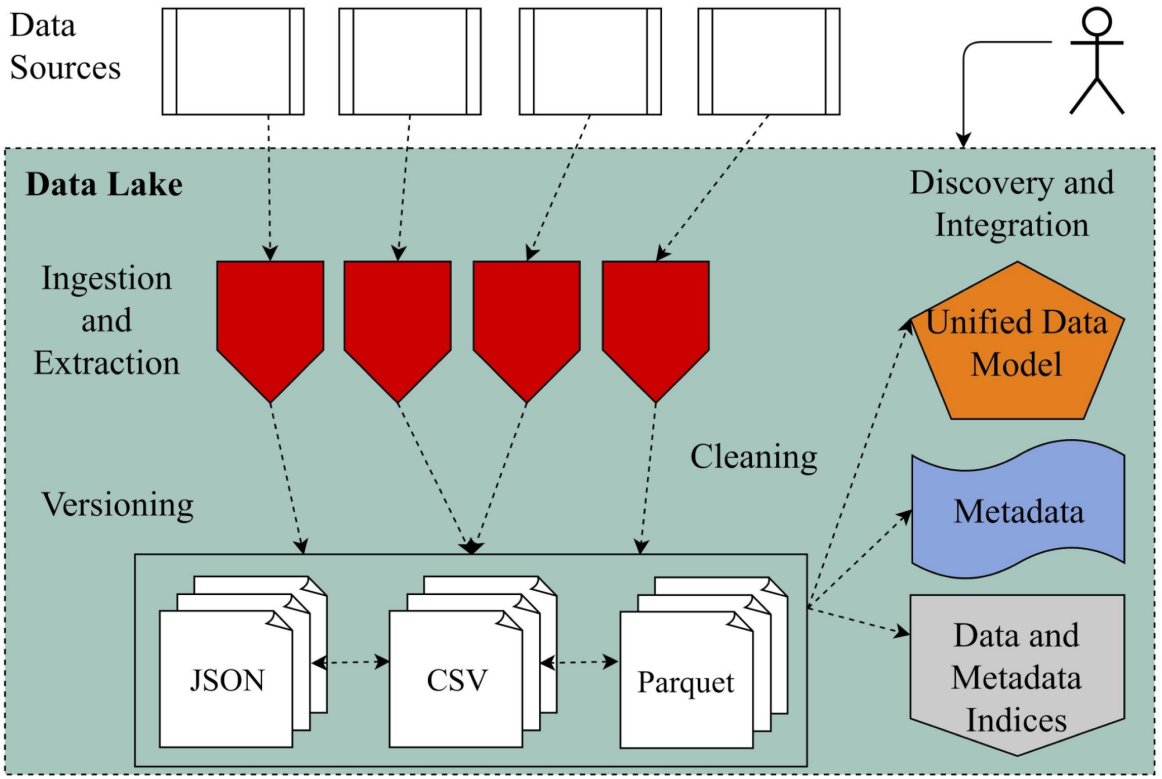
- Ingest already-structured data sets
- Extract structured data from unstructured and semi-structured data sources

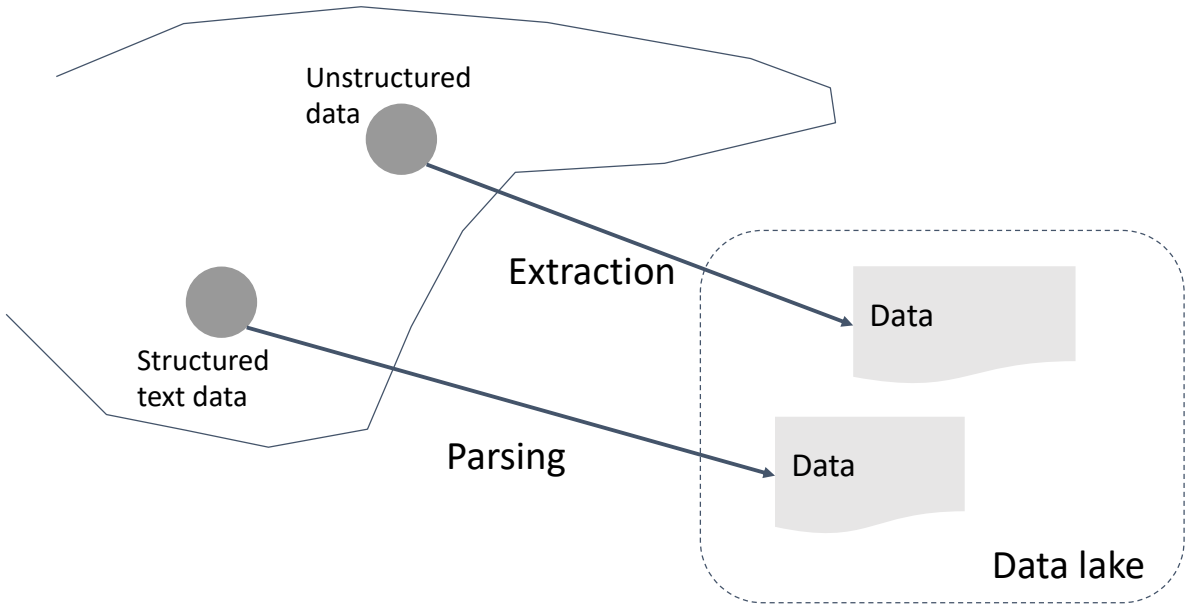
Data Integration is the task of finding joinable or union-able tables or of on-demand population a schema with all data from the lake that conforms to the schema

- Example: enrich Electronic Health Records (EPR) using data from various non-standard personal health record data sets for better predicting health risks
- Joining or “unioning” tables from different data sets and sources

Common Tasks in Building Data Lakes

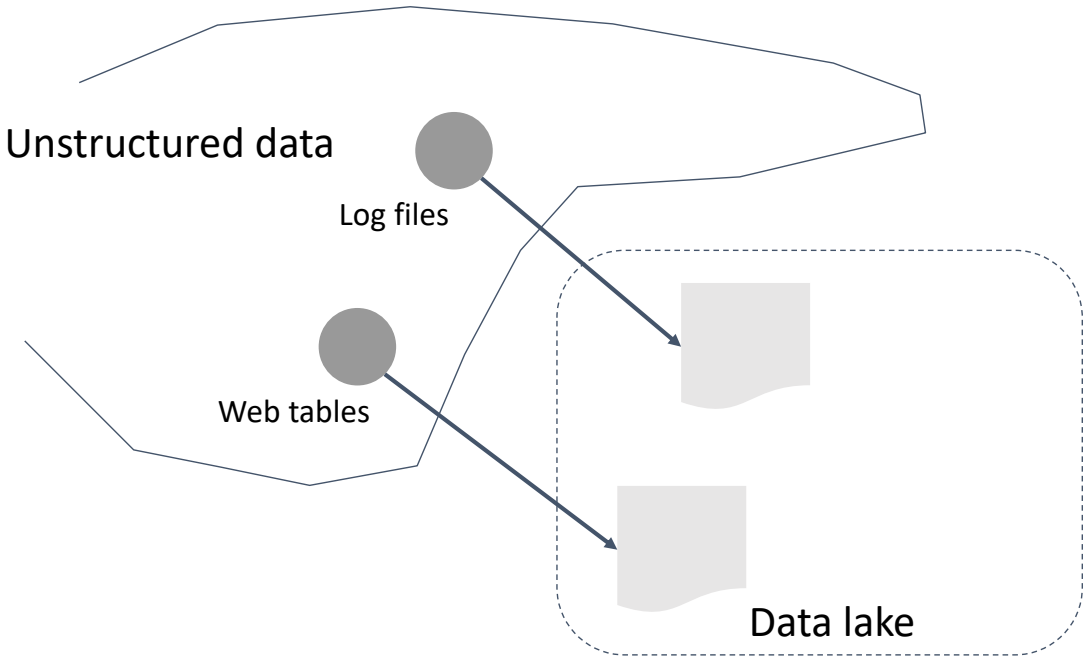
- Ingestion
- Extraction (Type Inference)
- Metadata Management
- Cleaning
- Integration
- Discovery
- Versioning



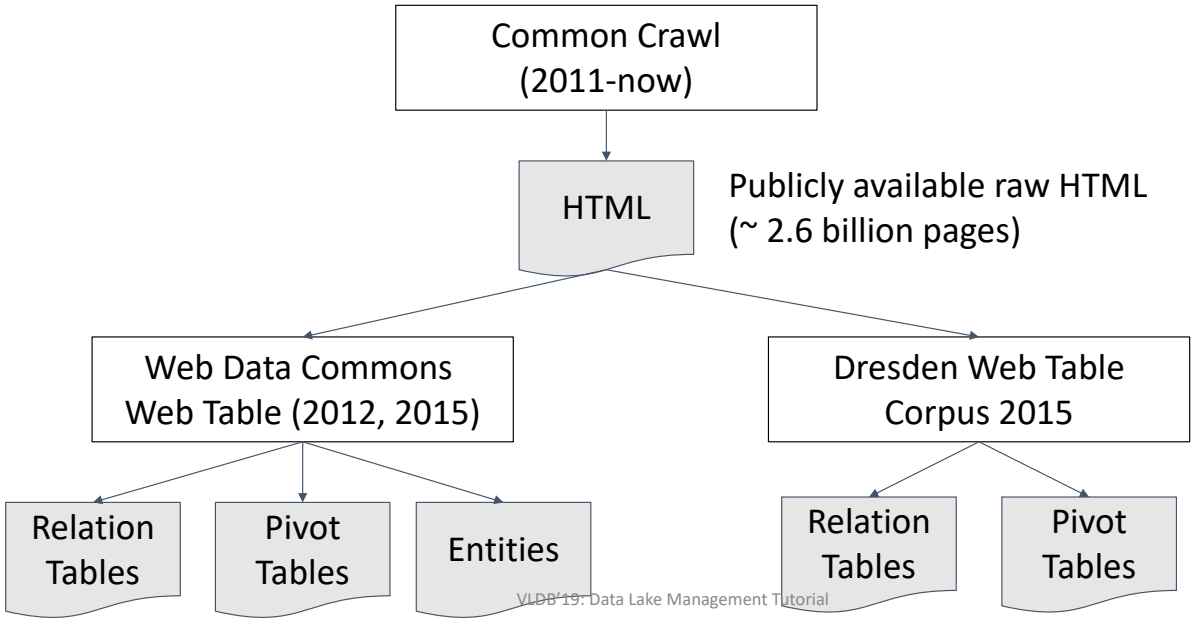

Unstructured
data
Extraction
Data
Structured
text data
Parsing
Data
Data lake

Ingestion

Extraction



Public Access of Web Tables



WDC Web Tables <http://webdatacommons.org/webtables/>

233 million Web tables (2015)

- 90 million relational tables
- 140 million entities
- 3 million matrices (pivot tables)

Corpus also provides metadata containing

- Orientation of records (row vs column)
- Title of HTML page
- Text surrounding the extracted table

Dresden Web Table Corpus

125 million tables, where the data table classifier is learned from a training set

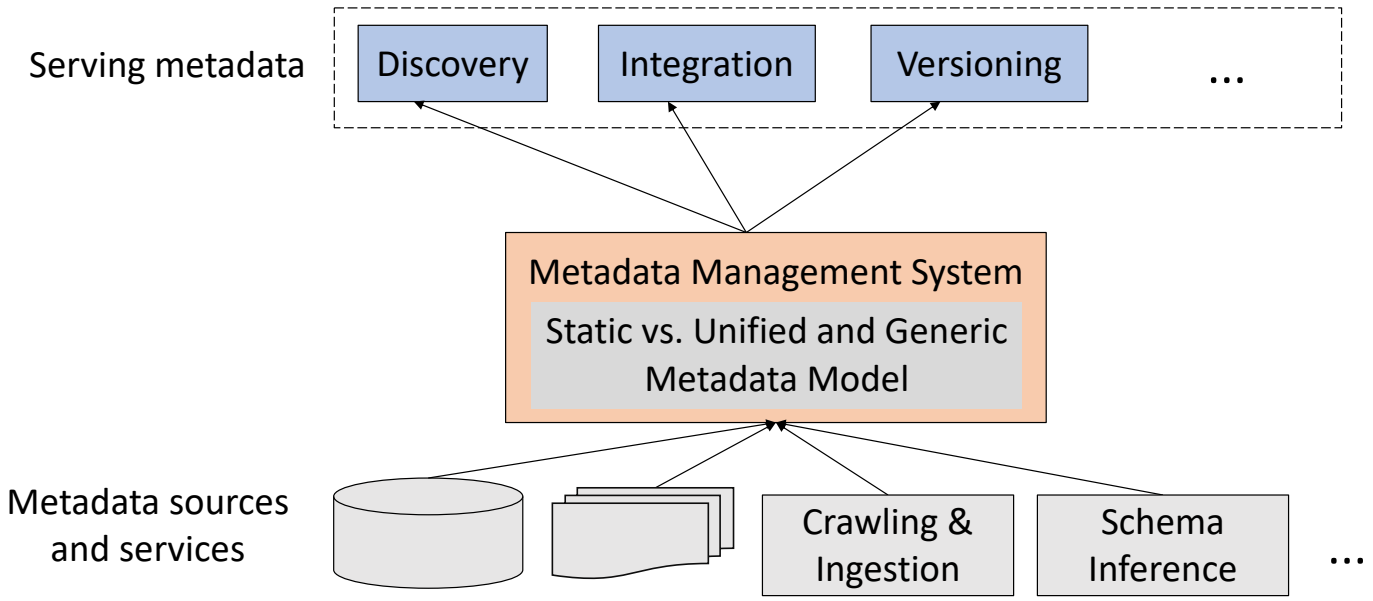
- Horizontal, vertical and matrix (pivot) tables

The classifier

- 127 features are generated for each HTML <table>
- Max and average cell count per row and column
- Max and average cell length in characters
- Frequency of cells containing <TH>, <A>, <IMG>, …

The training data with 2000 data tables is manually curated

Correlation based feature selection (CFS) is used to perform feature selection, producing 30 useful features



Architecture of Meta Data Management

Metadata Management in Data Lakes: Ideas

Enrich data and meta data with semantic information, support template-based queries on metadata

Extract deeply embedded metadata and contextual metadata to support topic-based discovery

Integrate metadata about data, users, and queries, and support visual analytics

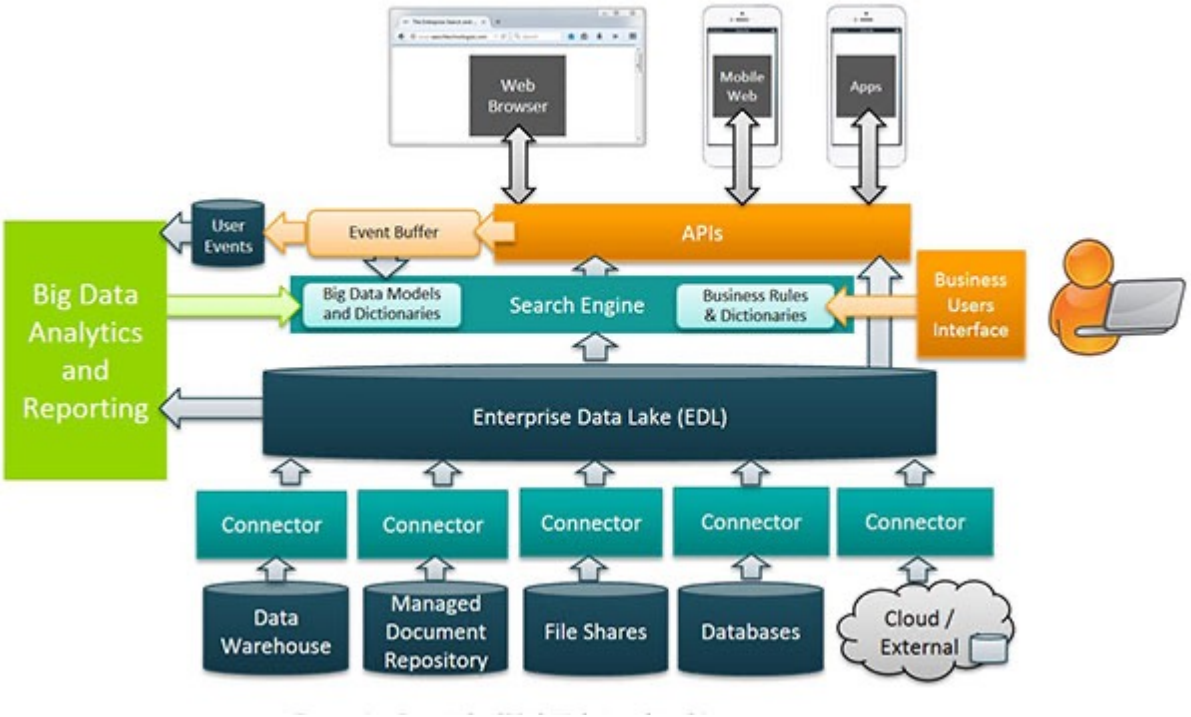
GOODS: Enterprise Metadata Model

Google internal data lake has tens of billions of data sets, some of giga- and tera-bytes

For each data set, an entry contains basic metadata, provenance, schema, content summary, …

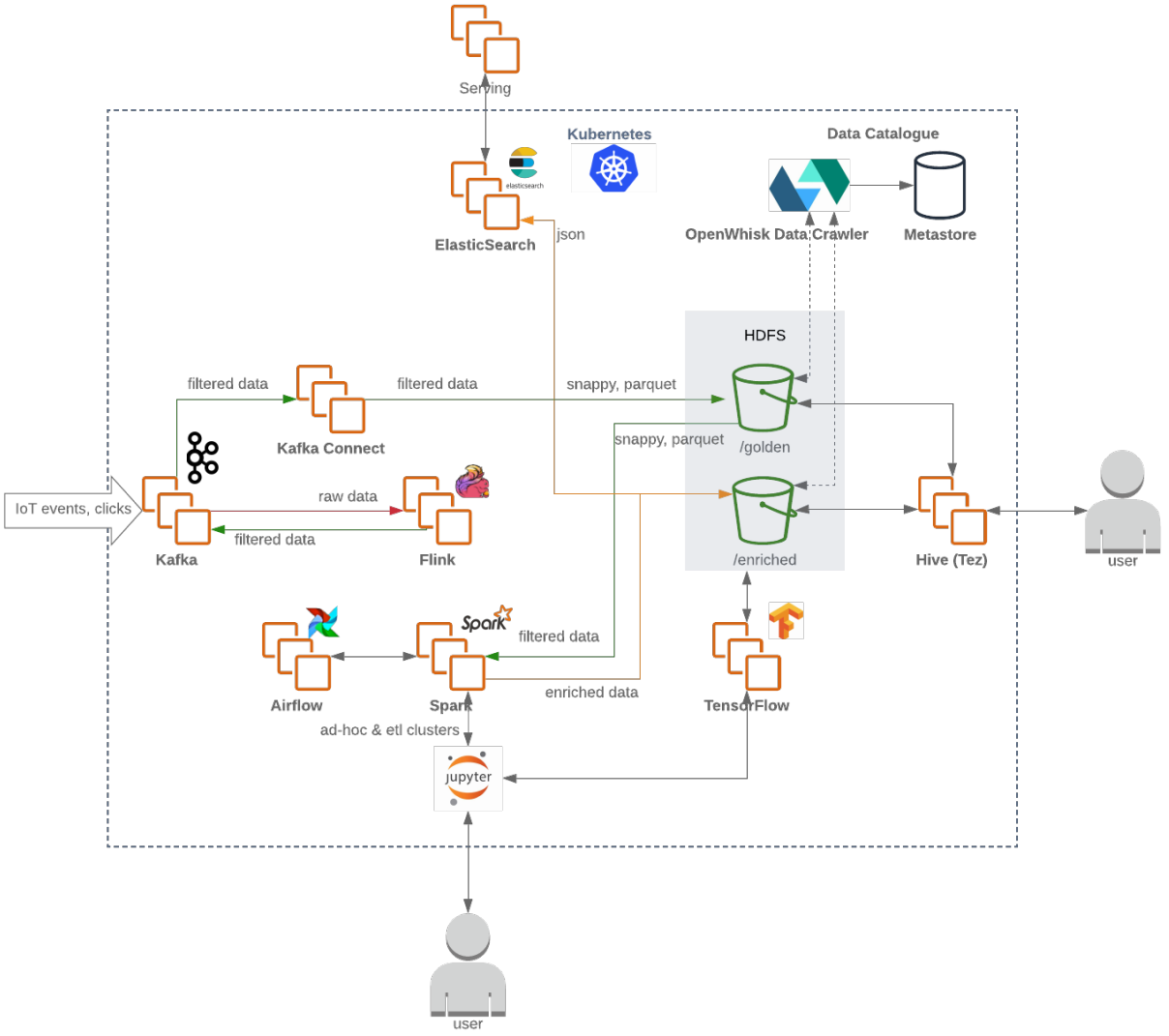
Sources of metadata: content samples, logs, source code repositories, crowdsourcing, knowledge bases, …

Path/Identifier	Metadata					
	size	…	provenance	…	schema	…
/bigtable/foo/bar	100 G		written by job A		Proto:foo.bar	
/gfs/nlu/foo	10G		written by job B read by job C		Proto:nlu.foo	



Architecture

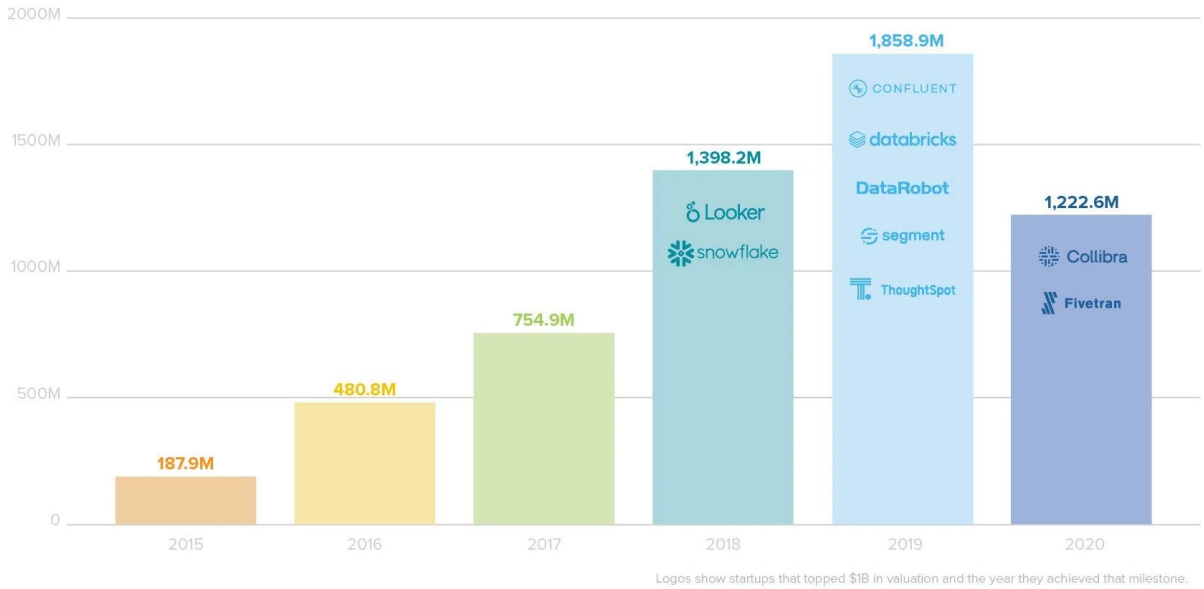
Using Open Source Systems



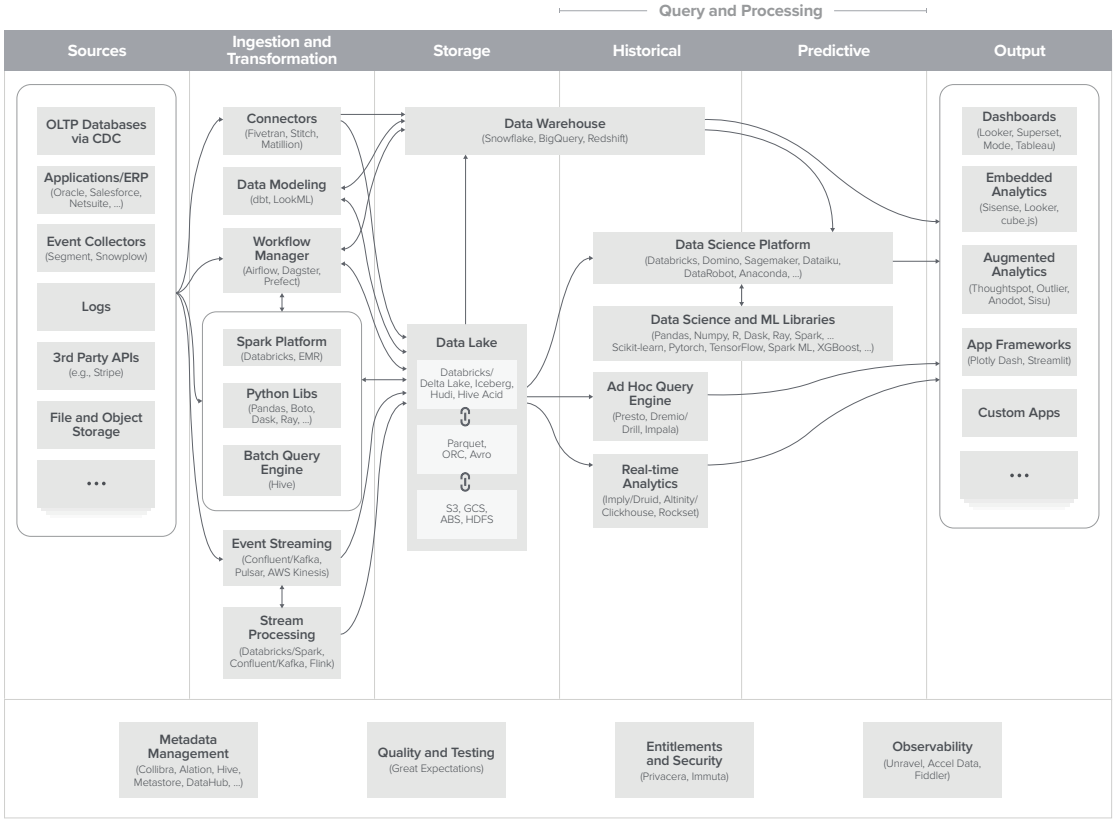
Enterprise Data Infrastructure

Analytics systems + operational systems

Spending on data infrastructure: $66 billion in 2019



A Unified Data Infrastructure Architecture



Interpreting the Architecture

Query and Processing

Sources	Ingestion and Transformation	Storage	Historical	Predictive	Output
Generate relevant business and operational data	Extract data from operational systems (E) Deliver to storage, aligning schemas between source and destination (L) Transform data to a structure ready for analysis (T)	Store data in a format accessible to query & processing systems Optimize for low cost, scalability, and analytic workloads (e.g., column store) In some cases, provide additional data structures or guarantees	Provide an interface for analysts and data scientists to derive insights (query) Execute queries and data models against stored data, often using distributed compute (processing) Describe what happened in the past (including very recent past)	Predict what will happen in the future Build data-driven/ ML applications	Present results of data analysis to internal and external users Embed data models into operational systems and applications

Coordinate the flow of data and the execution of computations across the full lifecycle

Ensure proper data quality, performance, and governance of all systems and datasets

https://a16z.com/2020/10/15/the-emerging-architectures-for-modern-data-infrastructure/

Two Purposes and Ecosystems

Analytic use cases: decision support based on data

- Data warehouses
- Structured data, SQL/Python

Operational use cases: data intelligence in applications

- Data lakes
- Raw data, Java/Scala, Python, R, and SQL

Potential convergence of data warehouses and data lakes

Lakehouses

Transaction support

Schema enforcement and governance

BI support

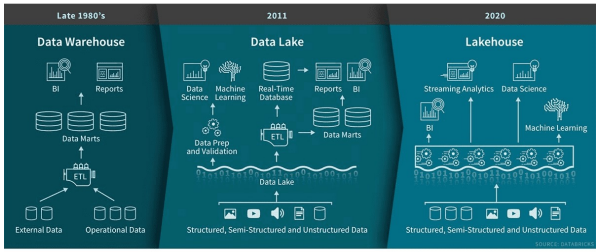
Decoupling between storage and compute

Open storage formats

Support for diverse data types, ranging from unstructured to structured

Support for diverse workloads: data science, machine learning, SQL and analytics

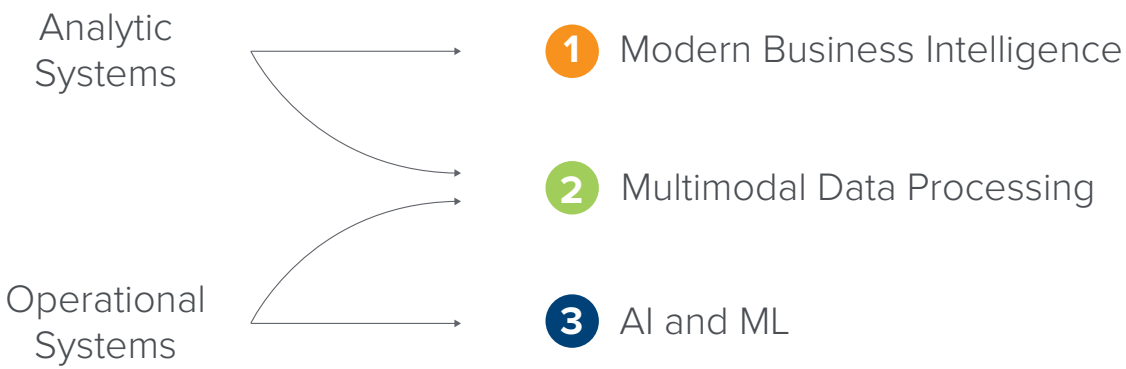
End-to-end streaming: real-time reports



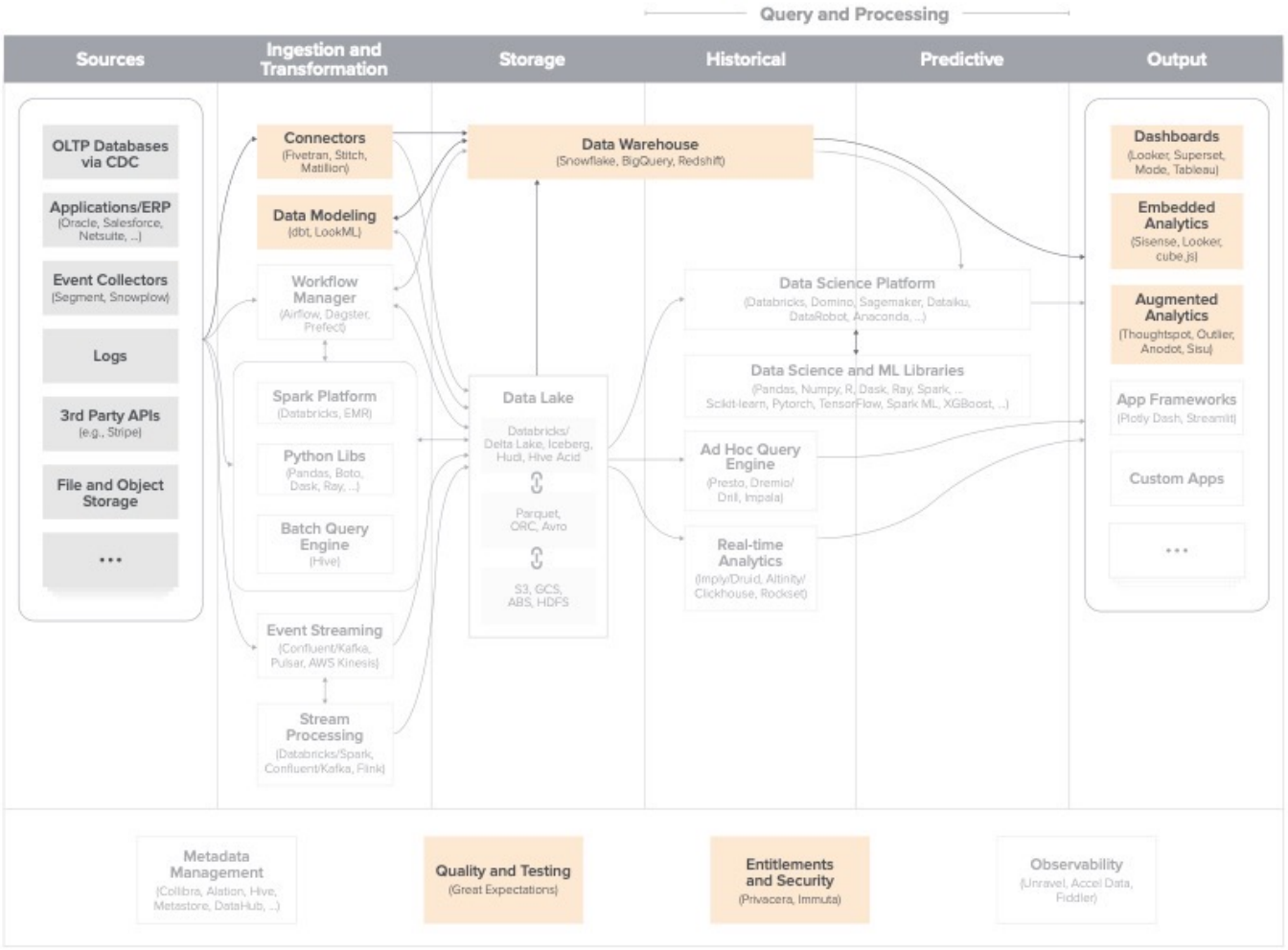
Architecture Shifts in Data Infrastructure

On Prem → Cloud Data Warehouse	Data warehouses are moving to the cloud with increased flexibility, scale, and ease of use—allowing any company to be a data company	snowflake, Google Big Query
Hadoop → Next-gen Data Lakes	Data lakes and related systems are becoming more performant and reliable, adding RDBMS-like features including ACID transactions and interactive SQL queries	databricks, presto
ETL → ELT	Brittle ETL processes (extract-transform-load) are being replaced with more flexible and consistent ELT pipelines (extract-load-transform)	Fivetran, dbt
Workflow Manager → Dataflow Automation	Data flow automation systems are helping to orchestrate thousands of data pipelines with a cleaner abstraction and modern executor integrations	PREFECT, DAGSTER, Apache Airflow
Analyst Teams → Self-serve Insights	Reporting, dashboarding, and automated analysis tools are becoming more available to non-technical users	Looker, Superset
Endpoint Protection → Global Data Governance	Data security and privacy measures (e.g., access controls) are becoming centralized on the data platform as use of data is increasingly regulated and user endpoints are harder to protect	Collibra, PRIVACERA

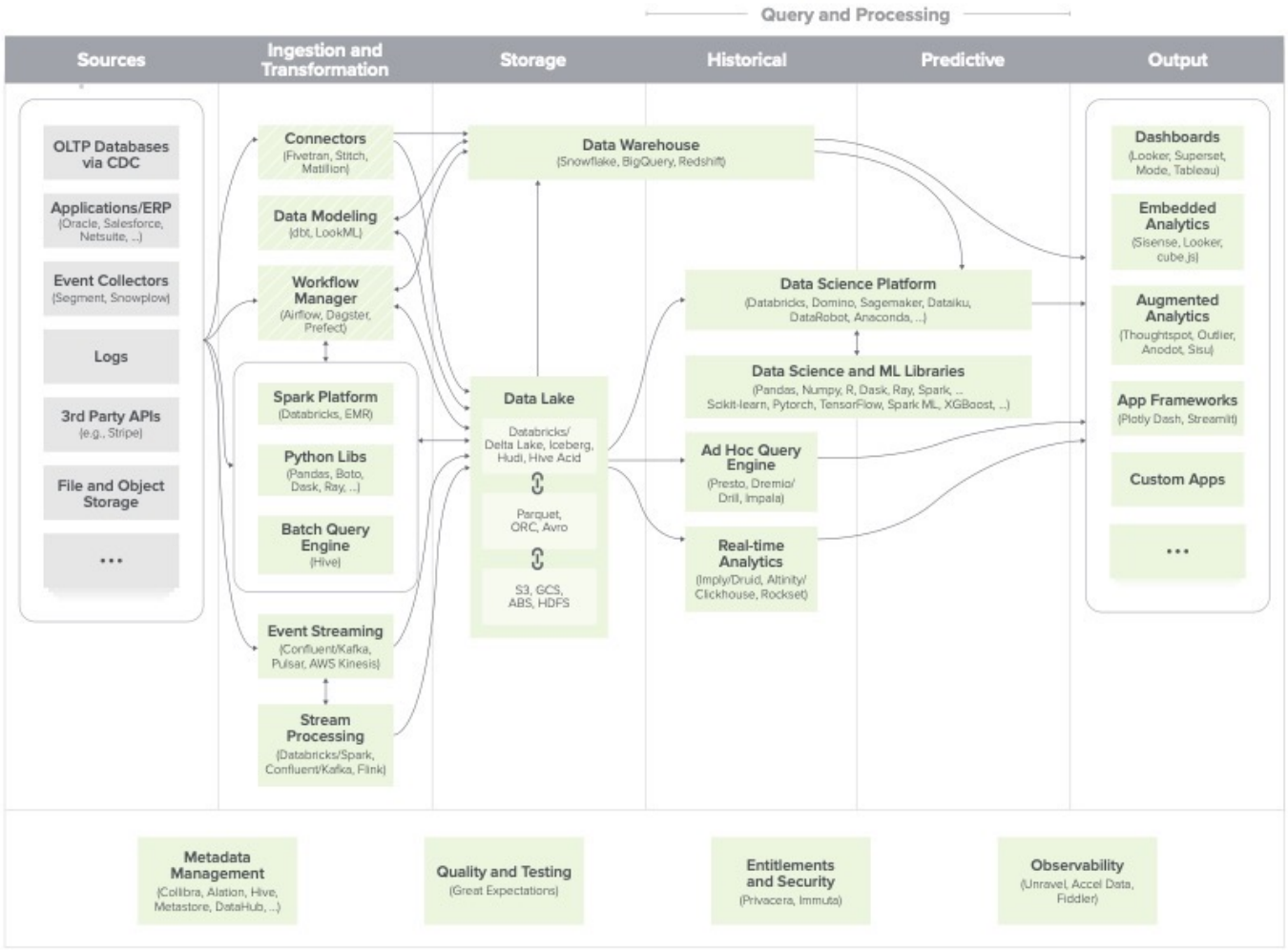
Three Common Blueprints



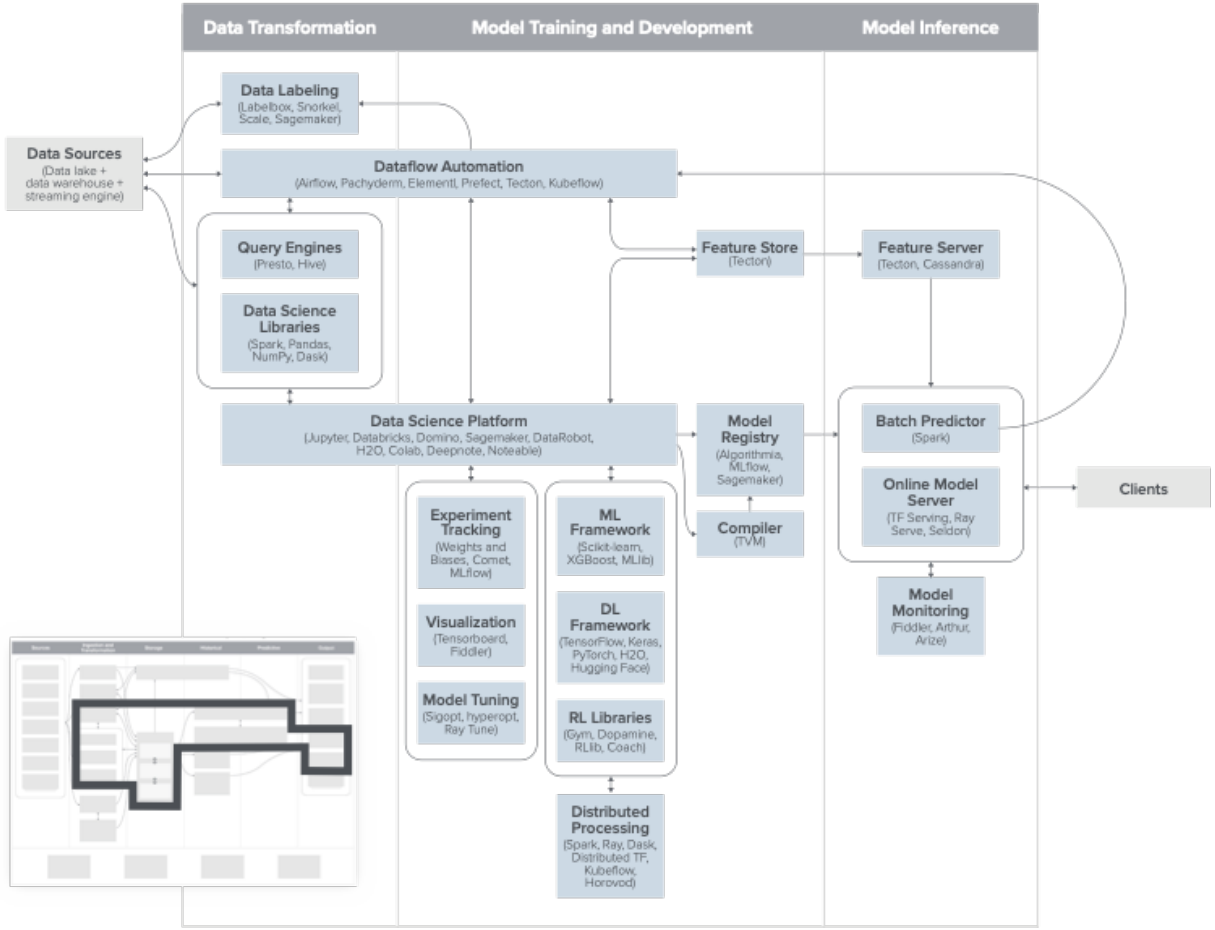
1. Modern Business Intelligence Blueprint



2. Multimodal Data Processing Blueprint



3. AI and ML Blueprint



Summary

Data lakes: motivations and concept

Major challenges and tasks in data lakes and enterprise data infrastructure

Data lake architecture

Enterprise data infrastructure