



SCHOOL *of* DATA SCIENCE

Data Lakes vs. Data Warehouses



Big Data Systems
School of Data Science
University of Virginia

Data Lake

Stores data with any structure such as:

- transactional data
- videos (unstructured)
- config files (semi-structured JSON)

Data Lake, contd.

Not necessary to define schema ... flexible

Amazon S3 is popular commercial product

- object file storage
- store data in buckets
- secure
- scales easily

Object File Storage

Data organized into objects

Each object contains:

- data
- metadata (key-value pairs describing object)
- unique ID (for retrieval)

Object File Storage, contd.

There are no actual folders (no directory structure)

Folders (e.g., in S3) are logical

Objects accessed via API or HTTP endpoints

Data Warehouse

Stores set of relational databases

Schema defined in advance ... not flexible

Examples: Amazon Redshift, Google BigQuery

Data Warehouse: GUI Example

The screenshot shows the Google Cloud Platform BigQuery interface. The top navigation bar includes 'Google Cloud Platform', 'Project for Coupler', a search bar 'Search products and resources', and user profile icons.

The left sidebar has a 'FEATURES & INFO' section, a 'SHORTCUT' button, and a 'HIDE PREVIEW FEATURES' link. It also displays pinned projects: 'project-for-coupler' (selected) containing 'Applicants' (selected), and 'feisty-audio-282807'.

The main area shows the 'Applicants' table preview. The table has columns: Row, id, Position, Application_Date, Stage_Name, Applicant_Status, Recruiter_Name, and Country. The data shows 10 rows of recruiter information from various countries like United Kingdom, Philippines, Colombia, Afghanistan, China, Russia, Mongolia, Belarus, and Norway.

Row	id	Position	Application_Date	Stage_Name	Applicant_Status	Recruiter_Name	Country
1	199	Recruiter	2019-10-07	RPI	lost	Howard Wolowitz	United Kingdom
2	211	Recruiter	2019-11-21	RPI	open	Leslie Winkle	Philippines
3	263	Recruiter	2020-02-04	RPI	won	Sheldon Cooper	Colombia
4	272	Recruiter	2020-04-02	RPI	lost	Raj Koothrappali	Afghanistan
5	323	Recruiter	2020-02-17	RPI	won	Howard Wolowitz	China
6	374	Recruiter	2019-10-04	RPI	lost	Leslie Winkle	Russia
7	376	Recruiter	2020-05-05	RPI	won	Leslie Winkle	Russia
8	389	Recruiter	2019-09-08	RPI	lost	Sheldon Cooper	Mongolia
9	401	Recruiter	2020-02-25	RPI	open	Leslie Winkle	Belarus
10	494	Recruiter	2020-04-09	RPI	lost	Howard Wolowitz	Norway

At the bottom, there are buttons for 'JOB HISTORY', 'QUERY HISTORY', and 'SAVED QUERIES'.

Comparison: Lake vs Warehouse

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	

Source: <https://aws.amazon.com/big-data/datalakes-and-analytics/what-is-a-data-lake/>