Cheatography

Data Warehousing - Data Lake Cheat Sheet

by [deleted] via cheatography.com/2754/cs/6550/

Data Lake

The term “data lake” has many definitions throughout the industry, ranging from a dumping ground for “to-be-used” data, to a more or less traditional EDW approach implemented on a big data platform. We would best define the data lake as an analytic system that allows data consolidation and analytic access with tunable governance. The data lake consists of a distributed, scalable file system, such as HDFS (Hadoop File System) or Amazon S3, coupled with one or many fit-for-purpose query and processing engines such as Apache Spark, Drill, Impala, and Presto..

Landing Area

This layer is where source data is stored in its full fidelity. This layer reduces barriers to onboarding new data, allowing early analytic access for new insights and the raw materials for “to-be” data products. Only very basic governance policies are required in the form of metadata (very often in the form of a partitioning schema) and information lifecycle management (security and disposition).

Data Lake

Data may be graduated from the landing area to the data lake. This data has basic governance policies, including data quality, retention, and metadata. It often has standard views or projections, allowing users to interact via familiar tools such as SQL, data exploration, and business intelligence tools.

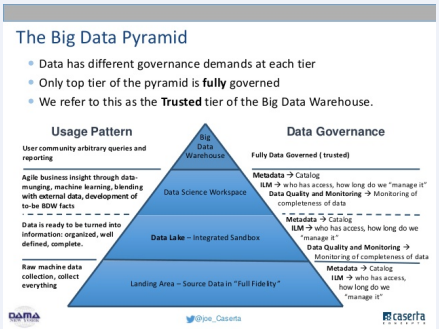
Data Science Workspace

This is the foundry of new data products. The work of data science may result in new data products, including new EDW facts.

Big Data Warehouse

This layer is fully governed, providing accurate, reliable, and well-d-efined information to data consumers. This big data warehouse may be platformed alongside the broader data lake, or in combination with traditional relational or MPP database technology.

Big Data Pyramid



The big data pyramid illustrates the different layers of the lake and what they represent from a data consumption and governance view.

Warehouse vs Lake

DATA WAREHOUSE	vs.	DATA LAKE
structured, processed	DATA	structured / semi-structured / unstructured, raw
schema-on-write	PROCESSING	schema-on-read
expensive for large data volumes	STORAGE	designed for low-cost storage
less agile, fixed configuration	AGILITY	highly agile, configure and reconfigure as needed
mature	SECURITY	maturing
business professionals	USERS	data scientists et. al.