Data Lake

- Collection of raw data in native format
  - Each element has a unique identifier and metadata
  - For each business question, you can find the relevant data set to analyze it
- Originally based on Hadoop
  - Enterprise Hadoop

Advantages of a Data Lake

- Schema on read
  - Write the data as they are, read them according to a diagram (e.g. code of the Map function)
  - More flexibility, multiple views of the same data
- Multi-workload data processing
  - Different types of processing on the same data
  - Interactive, batch, real time
- Cost-effective data architecture
  - Excellent cost/performance and ROI ratio with SN cluster and open source technologies

Principles of a Data Lake

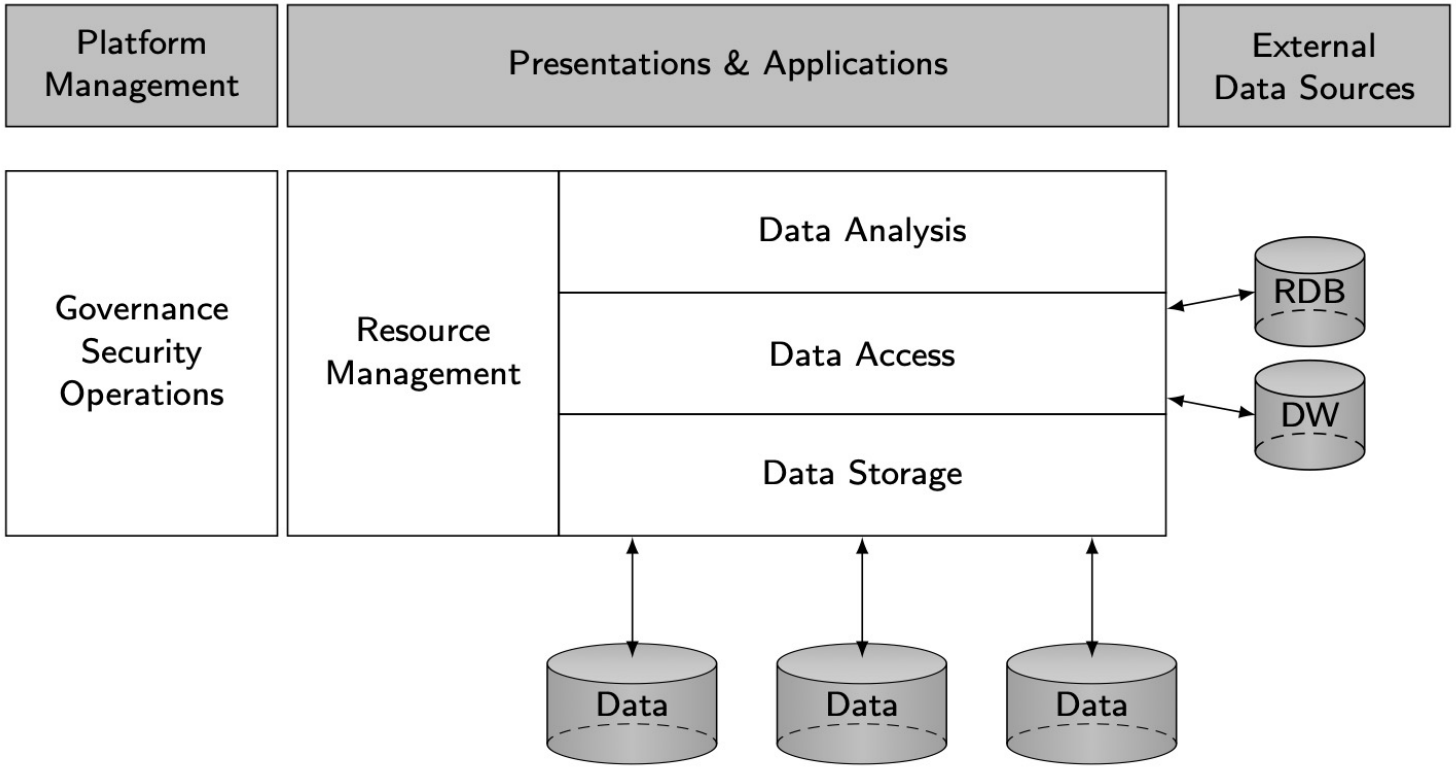
- Collect all useful data
  - Raw data, transformed data
- Dive from anywhere
  - Users from different business units can explore and enrich the data
- Flexible access
  - Different access paths to shared infrastructure
    - Batch, interactive (OLAP and BI), real-time, search,.....

Main Functions

- Data management, to store and process large amounts of data
- Data access: interactive, batch, real time, streaming
- Governance: load data easily, and manage it **according to a policy implemented by the *data steward***
- Security: authentication, access control, data protection
- Platform management: provision, monitoring and scheduling of tasks (in a cluster)

Data Lake Architecture

A collection of multi-modal data stored in their raw formats



Data Lake vs Data Warehouse

Data Lake

- Shorter development process
- Schema-on-read
- Multiworkload processing
- Cost-effective architecture

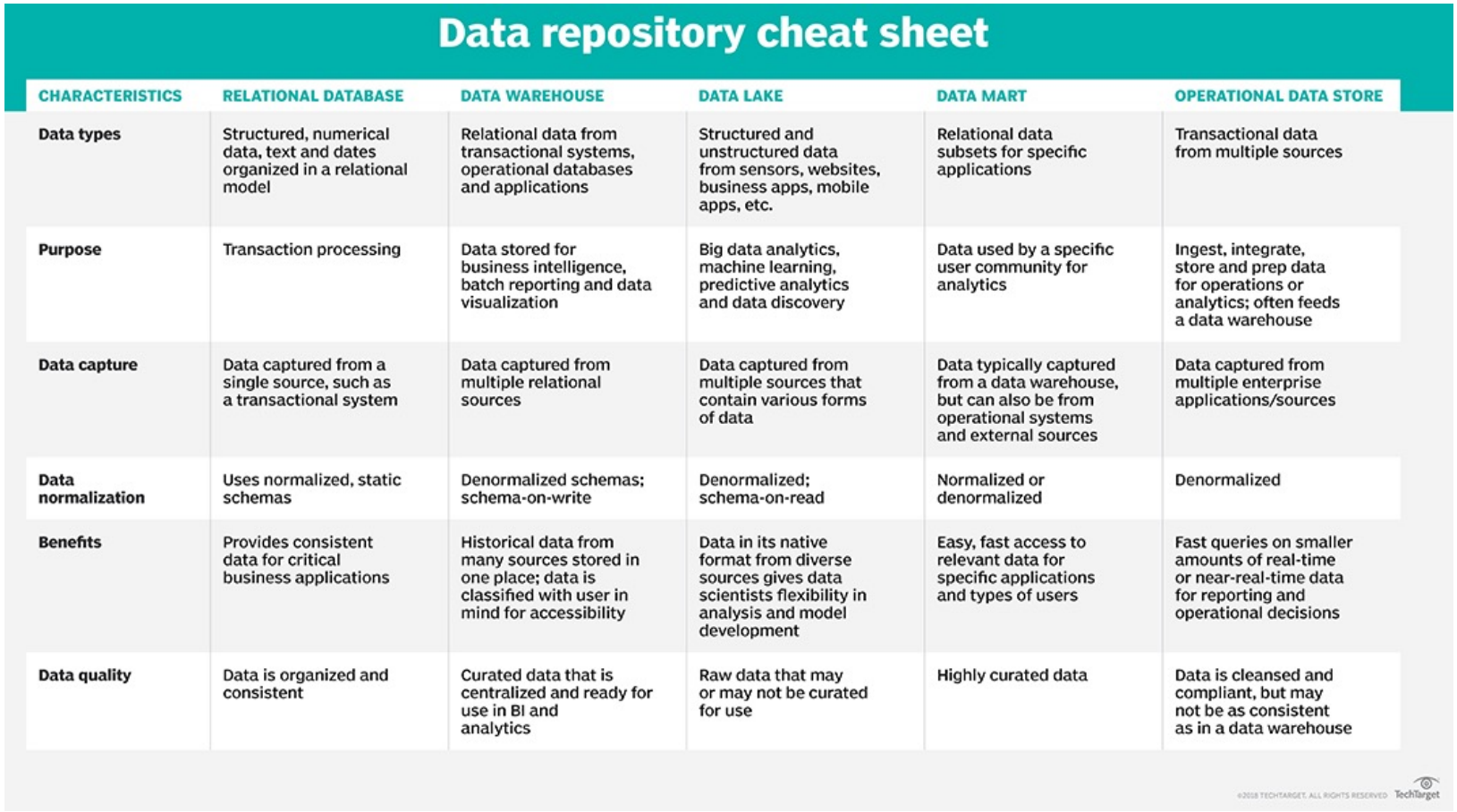
Data Warehouse

- Long development process
- Schema-on-write
- OLAP workloads
- Complex development with ETL

Data Lave Vs. Data Warehouse

Parameters	Data Lake	Data Warehouse
Storage	In the data lake, all data is kept irrespective of the source and its structure. Data is kept in its raw form. It is only transformed when it is ready to be used.	A data warehouse will consist of data that is extracted from transactional systems or data which consists of quantitative metrics with their attributes. The data is cleaned and transformed
History	Big data technologies used in data lakes is relatively new.	Data warehouse concept, unlike big data, had been used for decades.
Data Capturing	Captures all kinds of data and structures, semi-structured and unstructured in their original form from source systems.	Captures structured information and organizes them in schemas as defined for data warehouse purposes
Data Timeline	Data lakes can retain all data. This includes not only the data that is in use but also data that it might use in the future. Also, data is kept for all time, to go back in time and do an analysis.	In the data warehouse development process, significant time is spent on analyzing various data sources.
Users	Data lake is ideal for the users who indulge in deep analysis. Such users include data scientists who need advanced analytical tools with capabilities such as predictive modeling and statistical analysis.	The data warehouse is ideal for operational users because of being well structured, easy to use and understand.
Storage Costs	Data storing in big data technologies are relatively inexpensive then storing data in a data warehouse.	Storing data in Data warehouse is costlier and time-consuming.
Task	Data lakes can contain all data and data types; it empowers users to access data prior the process of transformed, cleansed and structured.	Data warehouses can provide insights into pre-defined questions for pre-defined data types.
Processing time	Data lakes empower users to access data before it has been transformed, cleansed and structured. Thus, it allows users to get to their result more quickly compares to the traditional data warehouse.	Data warehouses offer insights into pre-defined questions for pre-defined data types. So, any changes to the data warehouse needed more time.
Position of Schema	Typically, the schema is defined after data is stored. This offers high agility and ease of data capture but requires work at the end of the process	Typically schema is defined before data is stored. Requires work at the start of the process, but offers performance, security, and integration.
Data processing	Data Lakes use of the ELT (Extract Load Transform) process.	Data warehouse uses a traditional ETL (Extract Transform Load) process.
Complain	Data is kept in its raw form. It is only transformed when it is ready to be used.	The chief complaint against data warehouses is the inability, or the problem faced when trying to make change in in them.
Key Benefits	They integrate different types of data to come up with entirely new questions as these users not likely to use data warehouses because they may need to go beyond its capabilities.	Most users in an organization are operational. These type of users only care about reports and key performance metrics.

Data Repository Cheat-sheet



Data repository cheat sheet

CHARACTERISTICS	RELATIONAL DATABASE	DATA WAREHOUSE	DATA LAKE	DATA MART	OPERATIONAL DATA STORE
Data types	Structured, numerical data, text and dates organized in a relational model	Relational data from transactional systems, operational databases and applications	Structured and unstructured data from sensors, websites, business apps, mobile apps, etc.	Relational data subsets for specific applications	Transactional data from multiple sources
Purpose	Transaction processing	Data stored for business intelligence, batch reporting and data visualization	Big data analytics, machine learning, predictive analytics and data discovery	Data used by a specific user community for analytics	Ingest, integrate, store and prep data for operations or analytics; often feeds a data warehouse
Data capture	Data captured from a single source, such as a transactional system	Data captured from multiple relational sources	Data captured from multiple sources that contain various forms of data	Data typically captured from a data warehouse, but can also be from operational systems and external sources	Data captured from multiple enterprise applications/sources
Data normalization	Uses normalized, static schemas	Denormalized schemas; schema-on-write	Denormalized; schema-on-read	Normalized or denormalized	Denormalized
Benefits	Provides consistent data for critical business applications	Historical data from many sources stored in one place; data is classified with user in mind for accessibility	Data in its native format from diverse sources gives data scientists flexibility in analysis and model development	Easy, fast access to relevant data for specific applications and types of users	Fast queries on smaller amounts of real-time or near-real-time data for reporting and operational decisions
Data quality	Data is organized and consistent	Curated data that is centralized and ready for use in BI and analytics	Raw data that may or may not be curated for use	Highly curated data	Data is cleansed and compliant, but may not be as consistent as in a data warehouse