UNIVERSITÀ DEGLI STUDI FIRENZE
DINFO DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE
DISIT DISTRIBUTED SYSTEMS AND INTERNET TECHNOLOGIES LAB

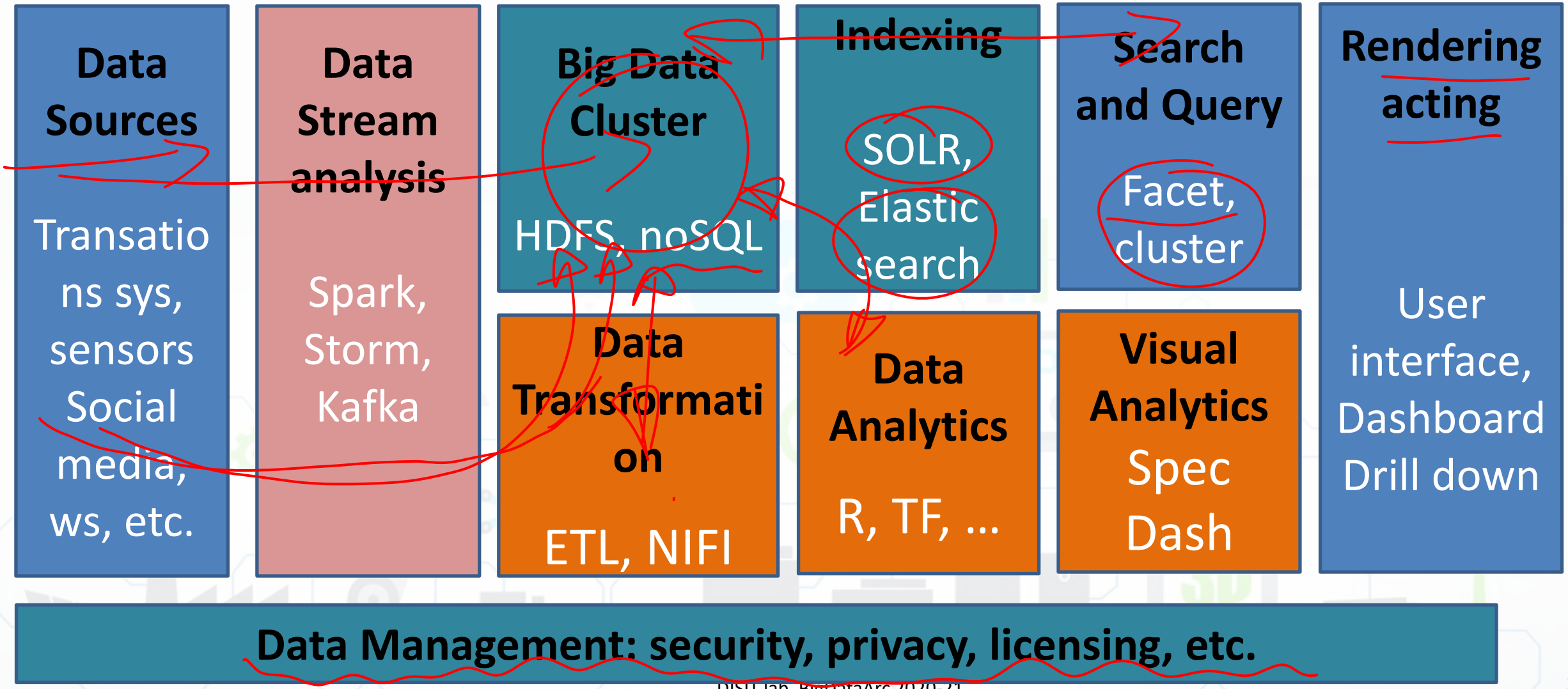
https://www.disit.org/
Paolo Nesi, paolo.nesi@unifi.it

Data Lake vs Data Warehouse

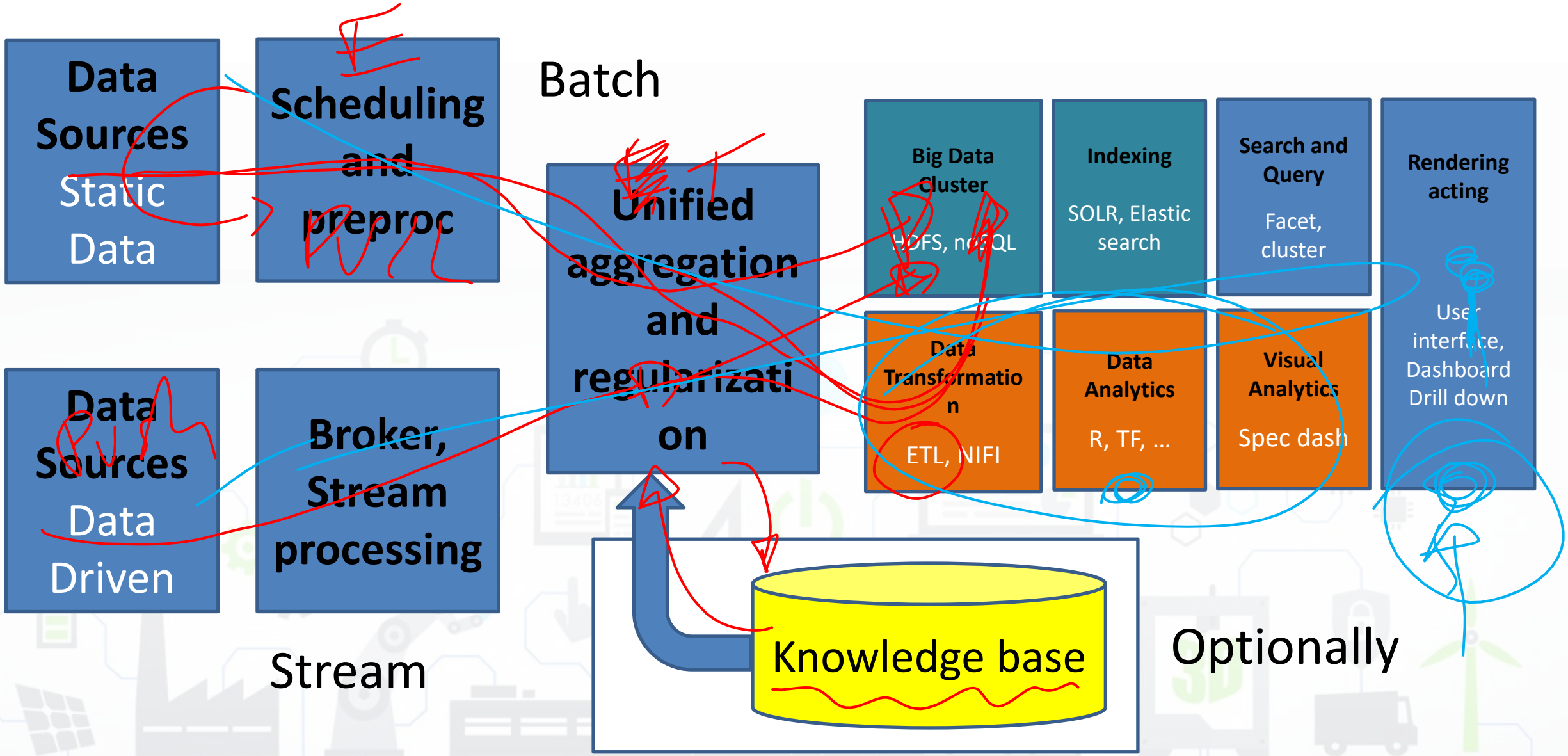
https://www.snap4City.org https://www.Km4City.org

Parte: x (2020)

Architettura di base Big Data, IOT, Industry 4.0



Lambda Architecture



Main Purpose

- To store a large amount of data, **big data**, and they can be **structured and un-structured**, several different kind of data:
  - **Direct Data**: Time series, geolocated data, events, shapes, measures, social media posts, video, files, logs, etc.
    - Most of them may have multiple features, e.g.: geolocated events with shape
  - **Derived Data**: predictions, typical trends, trajectories, flows, heatmaps, 3D reconstructions, traffic reconstruction, planning, simulations, etc.
- for **exploiting them** for producing:
  - Deductions, hints, early warning,
  - Derived Data as well, in real time

Decision Support

Main Functions

- **Data Extraction:**
  - gathering, harvesting, ingestion, reception in push, ...
- **Data Transformation:**
  - Adaptation, mapping, formatting, conversion
  - Cleaning or leaving as it is
- **Data Loading** and Refreshing: saving in the storage
  - As it is, converted and ready to use, etc.
- **Data Usage:**
  - As it is from Storage (faster, more rigid schema, higher volume in access)
  - Transformed on the fly (slower, more flexible, moderate volume in access)

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	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	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.

https://www.guru99.com/data-lake-vs-data-warehouse.html

Parameters	Data Lake	Data Warehouse
Task	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	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	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.

https://www.guru99.com/data-lake-vs-data-warehouse.html

Pros and Cons

- **Data Lake**
  - Original data preserved, structured and unstructured
  - Lower costs of ingestion
  - Lower performance in the usage
  - Security: possible control at source
  - More difficult to extend in usage, simpler in storage
  - More scientist oriented
    - Moderated results in access

- **Data Warehouse**
  - Original data transformed and prepared for mainly structured or semi-structured
  - Higher cost of ingestion
  - Higher performance in the usage
  - Security: Control on organized data
  - More difficult to extend in storage, simpler in usage
  - More Business and Purpose Oriented
    - Large volume of accesses

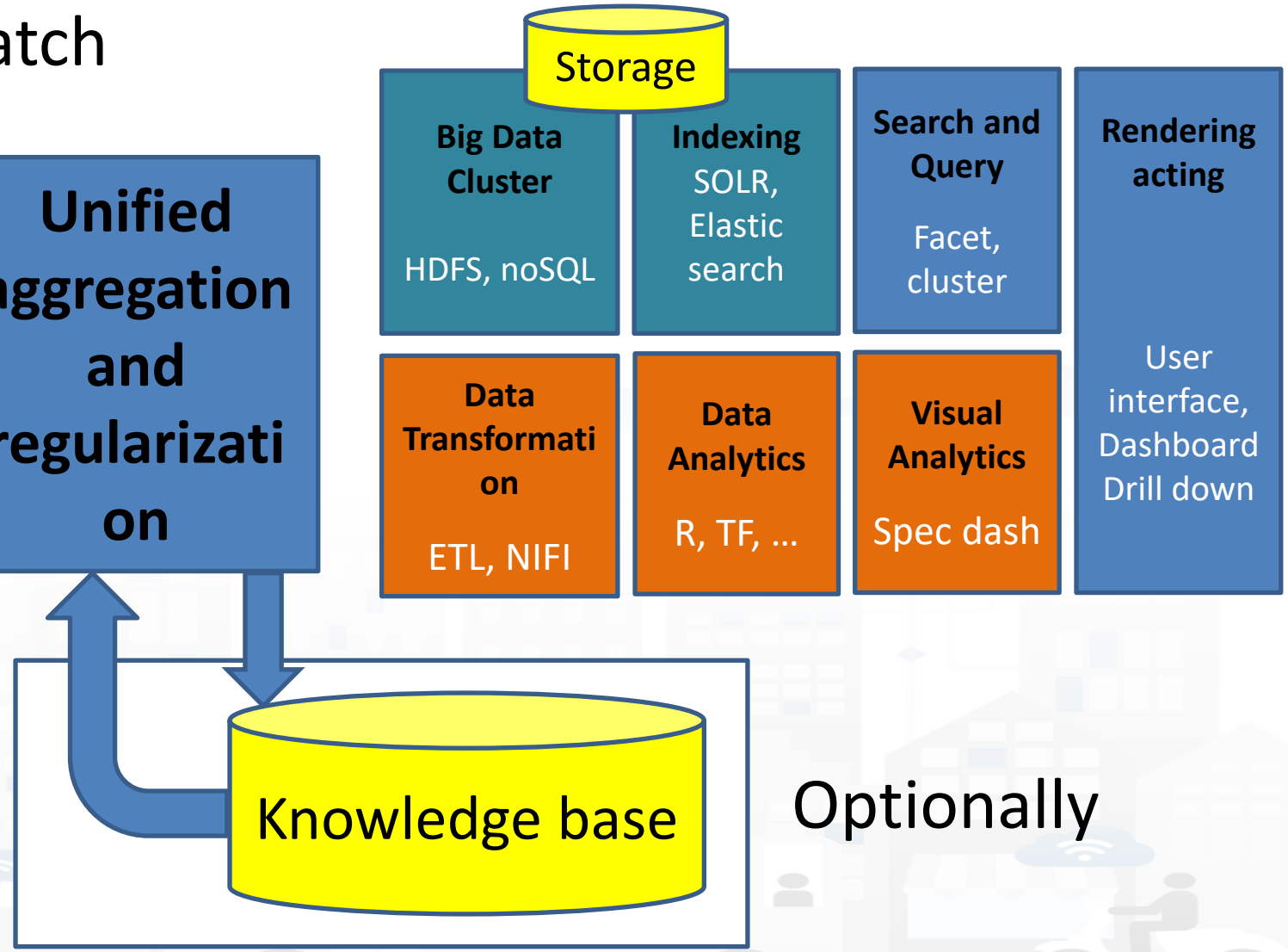
DWH

Data Sources Static Data

Scheduling and preproc

Batch

Unified aggregation and regularizati on



Data Sources Data Driven

Broker, Stream processing

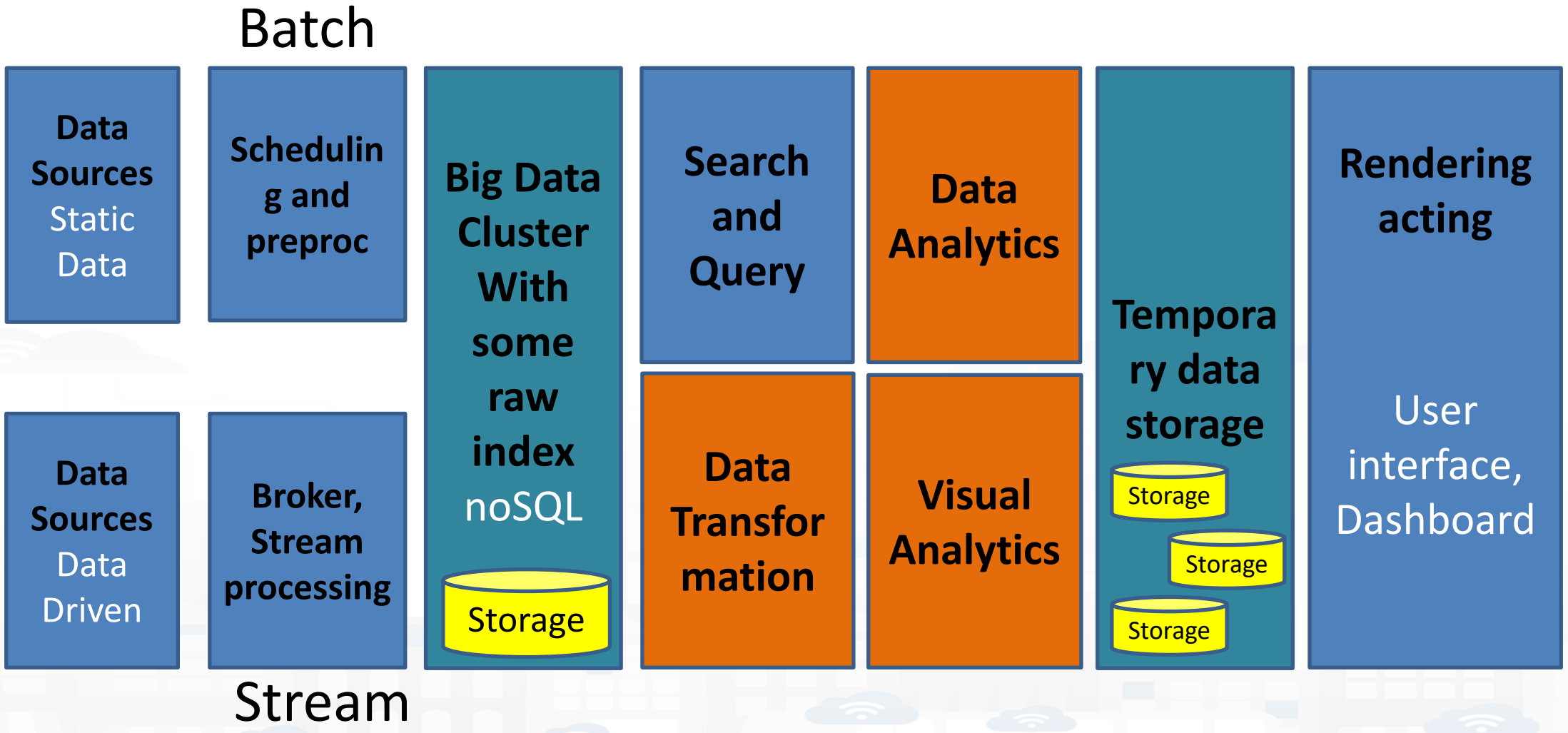
Stream

Optionally

Top Cloud Data Warehouses at a Glance

	Amazon Redshift	Microsoft Azure Synapse	Google BigQuery	Snowflake Cloud Data Platform
Initial Release	2012	2016	2010	2014
Separates Storage and Compute	No	Yes	Yes	Yes
Multi-Cloud	No	No	No	Yes
Query Language	Amazon Redshift SQL	TSQL	Standard SQL 2011 & BigQuery SQL	Snowflake SQL
Elasticity	Yes - Manual	Yes – Manual and Automatic	Yes – Automatic	Yes – Automatic
MPP	Yes	Yes	Yes	Yes
Columnar	Yes	Yes	Yes	Yes
Foreign Keys	Yes	Yes	No	Yes
Transaction	ACID	ACID	ACID	ACID
Concurrency	Yes	Yes	Yes	Yes
Durability	Yes	Yes	Yes	Yes
Automation	No	No	No	No
Website	Link	Link	Link	Link
Free Trial	Yes	Yes	Yes	Yes

DL


Batch
Data Sources Static Data
Scheduling and preproc
Big Data Cluster With some raw index noSQL
Storage
Search and Query
Data Analytics
Temporary data storage
Storage
Storage
Storage
Rendering acting
User interface, Dashboard
Data Sources Data Driven
Broker, Stream processing
Data Transformation
Visual Analytics
Stream

Combined Solutions