Logical architecture of a Big Data Platform

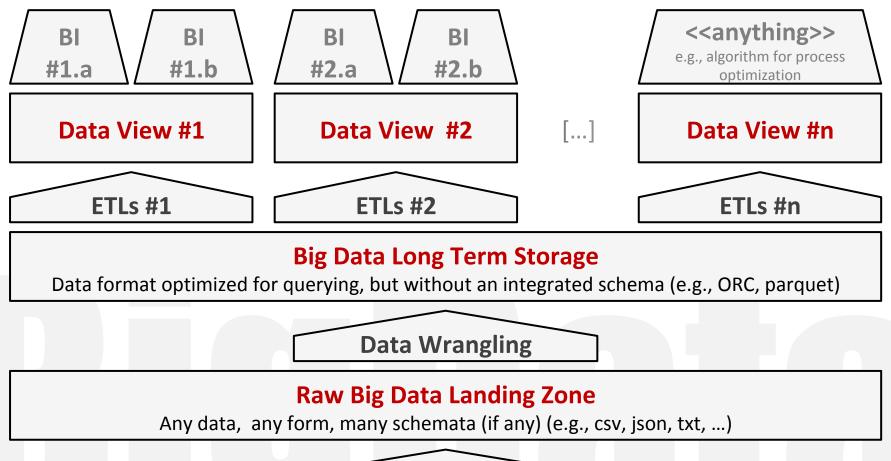
Emanuele Della Valle

DEIB - Politecnico di Milano

emanuele.dellavalle@polimi.it

http://emanueledellavalle.org

Logical Architecture of a Big Data Platform

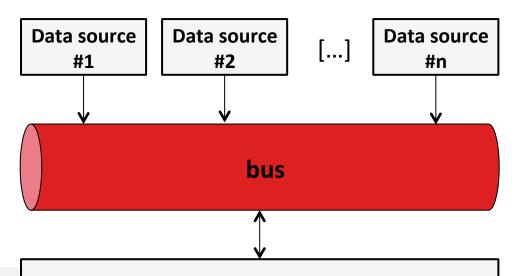


Data Ingestion

Data Ingestion

The process of importing, transferring and loading data for storage and later use

It involves loading data from a variety of sources It can involve altering and modification of individual files to fit into a format that optimizes the storage For instance, in Big Data small files are concatenated to form files of 100s of MBs and large files are broken down in files of 100s of MB.



Integration services

Transportation management, message routing, message brokering, transaction management, security, and transformation.

Data Wrangling

The process of cleansing "raw" data and transforming raw it into data that can be analysed to generate valid actionable insights. It includes understanding, cleansing, augmenting and shaping data. The results is data in the best format (e.g., columnar) for the analysis to perform.

Understand

What types of data are there, qualitative, quantitative, categorical? Is there any outliner?

Cleanse

remove outliners
adding missing
values
transform
quantitative values
in categorical ones

Augment

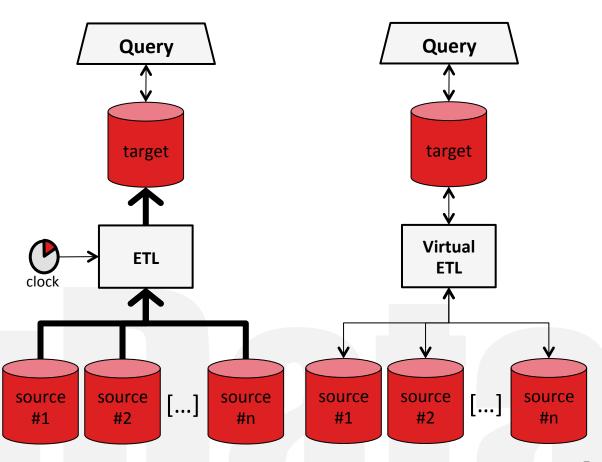
Aggregate data
within one data
source
Add data from
other data sources
based on exact/
fuzzy matches

Shape

Give data the best format for the analysis to perform

Extract Load Transform (in the Big Data era)

The process that extracts data from heterogeneous data sources, transform it in the schema that better fits the analysis to perform and load it in the system that will perform the analysis.



Thank you! Any Question?

Emanuele Della Valle

@manudellavalle

http://emanueledellavalle.org