

CLAIMS

1. A computing device comprising:
 - a processor;
 - a storage device coupled to the processor;
 - an engine stored in the storage device, wherein an execution of the engine by the processor configures the computing device to perform acts comprising:
 - receiving a raw dataset;
 - receiving one or more data quality metric goals corresponding to the received raw dataset;
 - determining a schema of the dataset;
 - identifying an initial set of validation nodes based on the schema of the dataset;
 - executing the initial set of validation nodes;
 - iteratively expanding and executing a next set of validation nodes based on the schema of the dataset until a termination criterion is reached; and
 - providing a corrected dataset of the raw dataset based on the iterative execution of the initial and next set of validation nodes.
2. The device of claim 1, wherein each validation node includes a data quality check and one or more remediation actions.
3. The device of claim 1, wherein execution of a validation node of the initial set of validation nodes comprises:
 - identifying all possible remediation actions for any data quality check;
 - transforming the data with each possible remediation action; and
 - computing a plurality of data quality metrics (DQMs) to evaluate the transformations.
4. The device of claim 1, wherein execution of a validation node includes a first stage, comprising:
 - performing a logical check of the raw dataset by a validator object to detect one or more anomalies in the raw dataset; and
 - performing different data transformations by way of a corresponding operator on the raw dataset to produce a new version of data for each data transformation, to correct the one or more detected anomalies.
5. The device of claim 4, wherein execution of the validation node includes a second stage comprising:
 - for each new version of data produced, generating a data quality metric (DQM) by an internal quality evaluator (IQE) module; and
 - generating a DQM for the raw dataset.

6. The device of claim 5, wherein each DQM of the second stage comprises at least one of (i) a summary of characteristics in multiple dimensions of the corresponding new version of data produced from the raw dataset; or (ii) a gain or change information of the corresponding new version of data produced from the raw dataset.
7. The device of claim 5, wherein execution of the validation node includes a third stage comprising:
for each DQM of each new version of data produced and the DQM of the raw dataset, performing a comparison to the raw dataset to assess an improvement from the raw dataset.
8. The device of claim 7, wherein execution of the validation node includes a fourth stage comprising:
selecting the operator of the new version of data produced that best meets the data quality metric goals.
9. The device of claim 8, wherein the operator that is selected has a highest gap between its corresponding DQM and the DQM of the raw dataset that is below a predetermined threshold.
10. The device of claim 1, wherein expanding a next set of validation nodes comprises at least one of:
determining a validation node that best achieves one or more received quality metric goals; or
determining a validation node based on mining an execution information repository to find all validation nodes that usually occur together.
11. A computer readable storage medium tangibly embodying a computer readable program code having computer readable instructions that, when executed, causes a computer device to carry out a method of improving data quality to conserve computational resources, the method comprising
receiving a raw dataset;
receiving one or more data quality metric goals corresponding to the received raw dataset;
determining a schema of the dataset;
identifying an initial set of validation nodes based on the schema of the dataset;
executing the initial set of validation nodes;
iteratively expanding and executing a next set of validation nodes based on the schema of the dataset until a termination criterion is reached; and
providing a corrected dataset of the raw dataset based on the iterative execution of the initial and next set of validation nodes.
12. The computer readable storage medium of claim 11, wherein each validation node includes a data quality check and one or more remediation actions.

13. The computer readable storage medium of claim 11, wherein execution of a validation node of the initial set of validation nodes comprises:
- identifying all possible remediation actions for each data quality check;
 - transforming the data with each possible remediation action; and
 - computing a plurality of data quality metrics (DQMs) to evaluate the transformations.
14. The computer readable storage medium of claim 11, wherein execution of a validation node includes:
- a first stage, comprising:
 - performing a logical check of the raw dataset by a validator object to detect one or more anomalies in the raw dataset; and
 - performing different data transformations by way of a corresponding operator on the raw dataset to produce a new version of data for each data transformation, to correct the one or more detected anomalies;
 - a second stage, comprising:
 - for each new version of data produced, generating a data quality metric (DQM) by an internal quality evaluator (IQE) module; and
 - generating a DQM for the raw dataset;
 - a third stage, comprising:
 - for each DQM of each new version of data produced, performing a comparison to the raw dataset to assess an improvement from the raw dataset; and
 - a fourth stage, comprising:
 - selecting the operator of the new version of data produced that best meets the data quality metric goals.
15. The computer readable storage medium of claim 14, wherein the operator that is selected has a highest gap between its corresponding DQM and the DQM of the raw dataset that is below a predetermined threshold.
16. The computer readable storage medium of claim 11, wherein expanding a next set of validation nodes comprises at least one of:
- determining a validation node that best achieves one or more of the data quality metric goals; or
 - determining a validation node based on mining an execution information repository to find all validation nodes that usually occur together.
17. A system comprising:
- a schema reference module configured to receive a raw dataset and to determine a schema of the raw dataset; and

a data synthesizer module coupled to the schema inference module and configured to receive one or more data quality metric goals corresponding to the received raw dataset from a knowledge base, wherein the data synthesizer module comprises:

an initializer module configured to identify an initial set of validation nodes based on the schema of the dataset;

an executor module coupled to the initializer module and configured to execute the initial set of validation nodes;

an expander module coupled to the executor module and configured to iteratively expand and execute a next set of validation nodes based on the schema of the dataset, until a termination criterion is reached; and

a terminator module coupled to the expander module and configured to iteratively determine the next set of validation nodes to consider by the expander module and to decide when to terminate the iterative determination; and

an information repository coupled to the executor and terminator modules and configured to provide a corrected dataset of the raw dataset based on the iterative execution of the initial and next set of validation nodes.

18. The system of claim 17, wherein each validation node includes a data quality check and one or more remediation actions.

19. The system of claim 17, wherein the initial set of validation nodes are configured to:

identify all possible remediation actions any data quality check;

transform the data with each possible remediation action; and

compute a plurality of data quality metrics to evaluate the transformations.

20. The system of claim 17, wherein each validation node comprises:

a first stage configured to:

perform a logical check of the raw dataset by a validator object to detect one or more anomalies in the raw dataset; and

perform different data transformations by way of a corresponding operator on the raw dataset to produce a new version of data for each data transformation, to correct the one or more detected anomalies;

a second stage configured to:

for each new version of data produced, generate a data quality metric (DQM) by an internal quality evaluator (IQE) module; and

generate a DQM for the raw dataset;

a third stage configured to:

for each DQM of each new version of data produced, perform a comparison to the raw dataset; and

a third stage configured to:

select the operator of the new version of data produced that best meets the data quality metric goals.

21. A computer implemented method of improving data quality to conserve computational resources, the method comprising
- receiving a raw dataset;
 - receiving one or more data quality metric goals corresponding to the received raw dataset;
 - determining a schema of the dataset;
 - identifying an initial set of validation nodes to be performed based on the schema of the dataset;
 - executing the initial set of validation nodes;
 - iteratively expanding and executing a next set of validation nodes based on the schema of the dataset until a termination criterion is reached; and
 - providing a corrected dataset of the raw dataset based on the iterative execution of the initial and next set of validation nodes.
22. The computer implemented method of claim 21, wherein each validation node includes a data quality check and one or more remediation actions.
23. The computer implemented method of claim 21, wherein execution of a validation node of the initial set of validation nodes comprises:
- identifying all possible remediation actions for each data quality check;
 - transforming the data with each possible remediation action; and
 - computing a plurality of data quality metrics (DQMs) to evaluate the transformations.
24. The computer implemented method of claim 21, wherein execution of a validation node includes:
- a first stage, comprising:
 - performing a logical check of the raw dataset by a validator object to detect one or more anomalies in the raw dataset; and
 - performing different data transformations by way of a corresponding operator on the raw dataset to produce a new version of data for each data transformation, to correct the one or more detected anomalies;
 - a second stage, comprising:
 - for each new version of data produced, generating a data quality metric (DQM) by an internal quality evaluator (IQE) module; and
 - generating a DQM for the raw dataset;
 - a third stage, comprising:
 - for each DQM of each new version of data produced, performing a comparison to the raw dataset; and
 - a fourth stage, comprising:

selecting the operator of the new version of data produced that best meets the data quality metric goals.

25. The computer implemented method of claim 21, wherein expanding a next set of validation nodes comprises at least one of:

determining a validation node that best achieves one or more received quality metric goals; or

determining a validation node based on mining an execution information repository to find all validation nodes that usually occur together.