**‘datacookie’ – automated data quality & preparation engine**

**Abstract:**

Data Quality serves as a critical success factor in not only data intensive projects but as an organisation’s data health. As the volume of data is growing exponentially, it becomes a tedious task to review and rectify data quality issues.

60% of the time spent by data scientists is in data cleaning and preparation. Enterprises are looking for advanced yet simpler ways to tackle the problem.

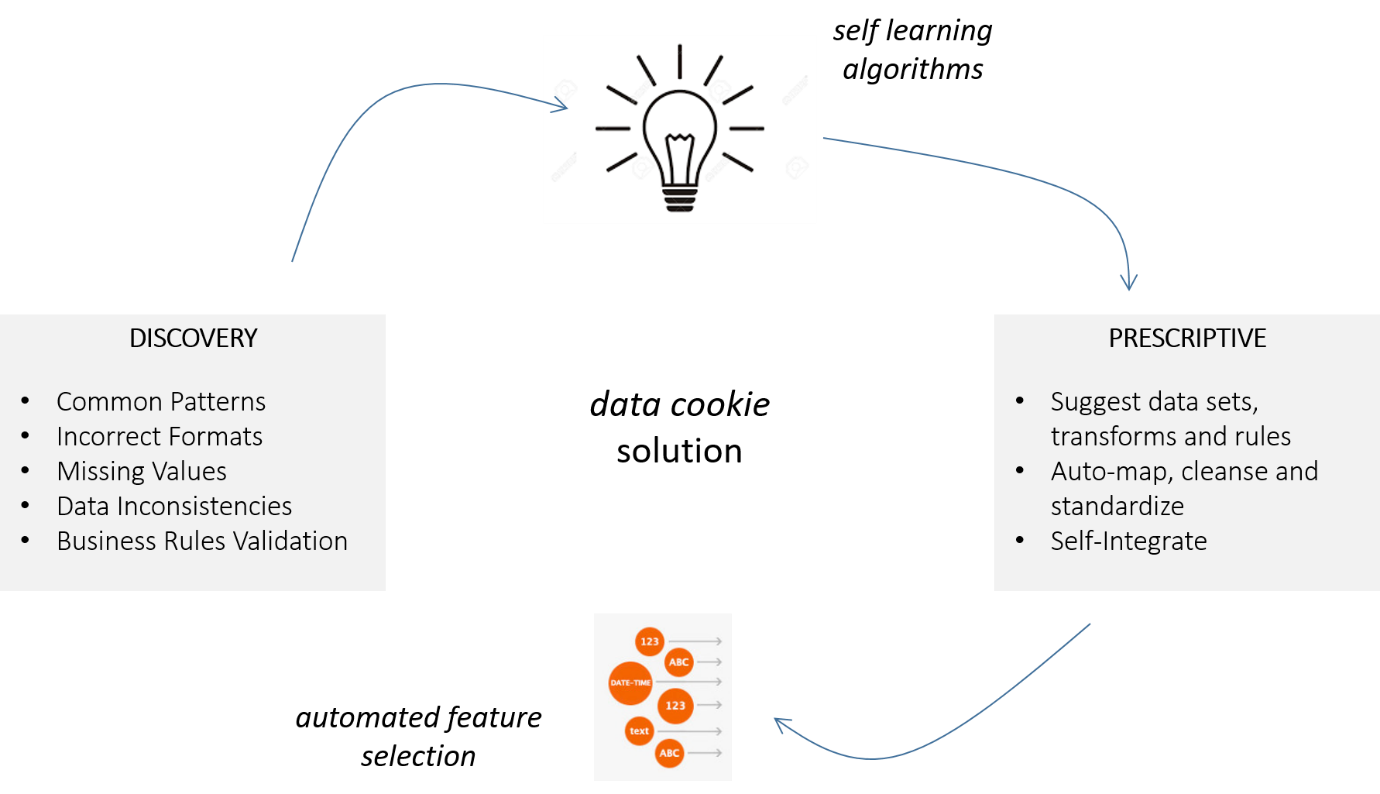
We propose an engine which will work to deliver an automated solution to identify data quality issues, enable on-the-fly data preparation and automatic feature selection to help both business users and data scientists to prepare data at the flick of a button.

This framework will be powered by machine learning to self-learn about the data to greatly enhance accuracy in identifying quality issues as well as empower itself to “suggest: corrections to dirty data. Currently, no existing tools like Paxata or Talend have a ML powered backend to serve the same purpose. Also, this tool will have the capability for various customizations as per client’s data requirements.

**Solution Approach:**

Data Quality is measured based on 3 fundamentals: Accuracy, Consistency and Completeness. The application is built upon the following building blocks:

1. Data Profiler – understands the semantics of the data – fields, attributes, etc
2. Business Rules Configurator – self-service utility which allows business user to apply their domain specific business rules
3. Scoring – assigns a qualitative score on the health of data
4. Tagged Data Archival – the historical repository from where the tool “learns”
5. Front End Utility – web based; allows business users to enable all actions – quality monitor, data preparation, automated feature selection.



The engine will focus on the following:

1. Data quality check and identifying the gaps based on generic rules as well as customised business standards based on an industry domain
2. Handling missing values with predictive algorithms, outlier treatment, unstructured data cleaning and variable suggestions for improving data preparation method
3. Automating feature selection using deep learning

**Design Architecture:**

“datacookie” loves big data. The system will be capable to process multitudes of rows in seconds and perform key data quality checks by leveraging in-memory computing and powerful caching. It will be deployed on a high performing cluster in the cloud to enable enterprise-wide access to multiple users.

SYSTEM FLOW:

User-inputted business rules and standard quality checks are applied on this data. The ‘data’ and ‘rule’ are tagged together and archived in a No-SQL repository. Advanced machine learning algorithms use this as train data to “self-learn” about the data being fed into this application. Users can create separate projects to let the system differentiate between various business data sets.

The next module spits out a detailed data quality score with directly pinpointing issues “row-wise”. The story doesn’t end here with the tool being capable of the following:

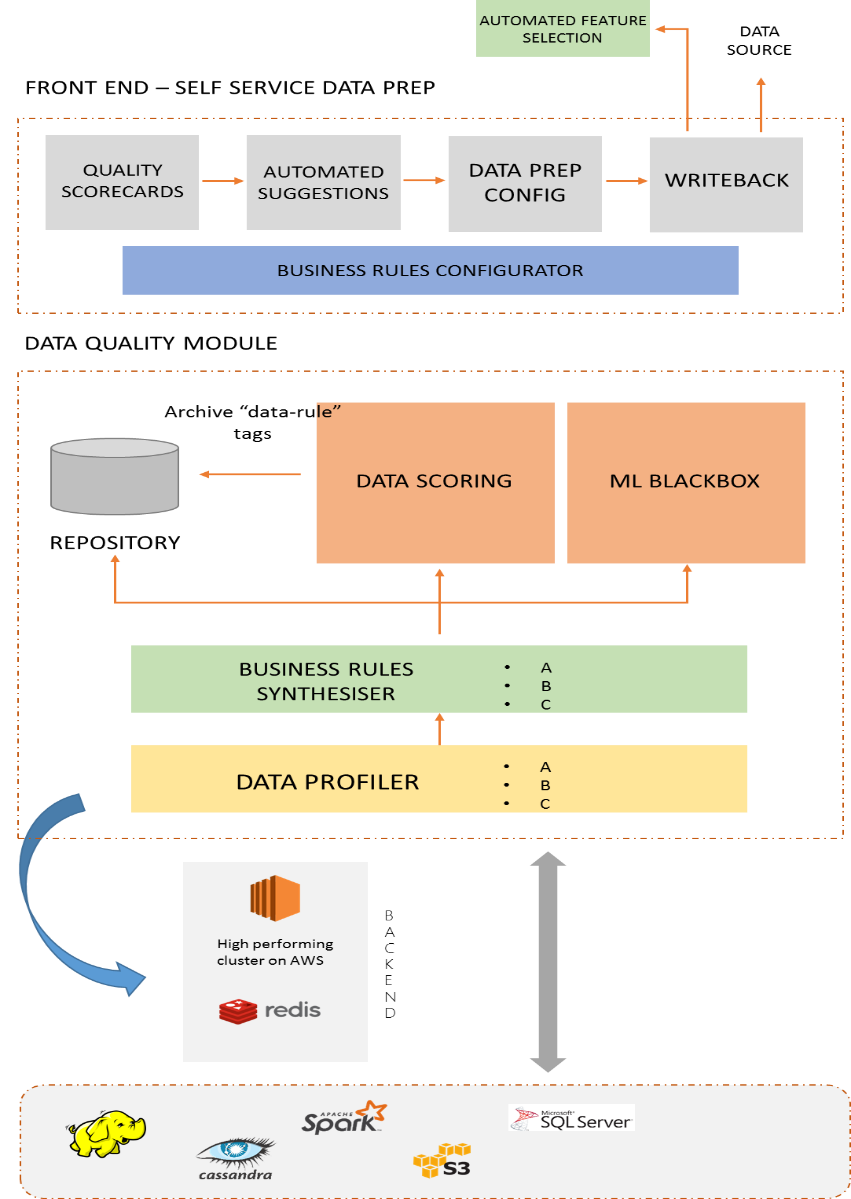
Auto-Mapping – detects master entities across master data model and applies quality rules

Self-Tune – based on historical information pinpoints outliers or erratic trends

Self-Secure – identifies sensitive data like social security number and offers to mask it

The user is offered suggestions to handle dirty data wherever identified. This paves the path for generic data preparation tasks also in-built in the system like.

The modifications are written back to the original data source using SQL connectivity or REST APIs.



**Data Quality Module**

1. **Data Profiler**: Data Profiler is a data profiling module that provides a fast and accurate way to assess the current condition of your data.
2. **Business Rules Synthesiser**: A comprehensive filter where one defines data quality rules, generic rules and domain specific rules.
3. **Data Scoring**: Data quality score will be produced on the basis of Business Rules Synthesiser and self-learning ML.

**Front End**

1. **Quality Scorecards**: An overall report of data on the basis Accuracy, Consistency & Completeness.
2. **Automated Suggestions & Data Prep Config**: Combining Business rules and quality scorecard, this will produce suggestions to prep data.
3. **Writeback**: Data will be send back to repository, from where it can be used further.

**Business Impact:**

60% of time data scientists spend cleaning data.

50% of time knowledge workers waste hunting & fixing data.

Improving data quality manually requires countless hours of data cleansing & preparation.

Getting data in the right shape and format can have a massive impact on the accuracy of a model (for the data scientist) and enterprise data health (for business user) and thus success of IT and business projects operations.

The market has no key product offering all apples in a singular basket: data quality management, data preparation and automated feature selection.

Incedo’s clients especially in the Finance and Pharma domains are data intensive and data rich organisations. In the era of big data, ‘datacookie’ aims to serve remedy to a chronic pain in the data lifecycle.

BENEFITS:

1. Enhanced Accuracy
2. Drastically reduced time
3. Reduced manual intervention – semi automation

~.~.~.~.~.~.~.~.~.~.~.~.~.~.~.~.~.~.~.~.~.~.~.~.~.~.~.~.