DQUEEN Develop Plan
Part: Meta

13. June, 2019

Dept. Biomedical informatics

Enzo Byun

- DQ Tool :: Meta part

1. DQ concept check list

1) 중복성

Eg. 테이블 내 데이터가 중복 되었는가?

환자의 처방 번호가 중복되었는가? 등

2) 완정성

Eg. 데이터 내 Missing 혹은 null 값이 존재하는가?

3) 논리적 타당성

Eg. 방문없이 발생한 처방 정보

4) 시간적 타당성

Eg. 입원시간은 퇴원시간보다 앞이어야함

5) 데이터 모델의 정합성 및 적합성

6) 데이터의 정확성

Eg. 사전에 정의된대로 들어 가있는가?

계산된 값이 정확한가?

7) 데이터의 일관성

Eg. 컬럼이 사전에 정의한대로 생성 되었는가 ?

데이터의 길이가 만족하는가?

- DQ Tool :: Meta part

2. 필요 기능
    - 사전 정의된 데이터 타입, 형식 일치 여부
    - 데이터의 중복 여부
    - Missing value 확인
    - 사전 정의된 데이터 모델 정의서와 일치 여부
    - DQ evaluation을 위한 rule 검사 수행
    - 수치 값들에 대한 Outlier 제공을 위한 Regression 등의 기능
    - Year, Monthly Trend 비교 기능
    - 범위 유효성 (Eg. 날짜 형식등) 검토
    - Data Quality Score를 위한 점수 계산
    - 데이터 Relation 검토

- DQ Tool :: Meta part

2. Tool Function definition

    - Count (Uniq Count, Count, Missing, null, 특수문자 …)
    - Compare with Pre definition Data model description (테이블 정의서)
    - Data model Relation Scan (Pk, Fk…)
    - Data type, data format Scan (생성된 데이터)
    - SQL Script execution function (Rule)
    - String value compare function
    - Time Trend and Conditional data ratio function
    - Continuous & Categorical variable outlier detection (linear regression, 3SD..)
    - Data Quality Score formula function
    - Visualization (hitmap, Matlab 3D plot, time period, cytoscape ..)
    - Text message alert
    - DB Volume check

- DQ Tool :: Meta part

3. system running process

- Pre computed -> Data evaluation -> Post computed -> Visualized -> Run dash board
  - Pre computed
    1) DB check : Table name, Column name
    2) Data type check
    3) DB volume, Table volume
    4) Row count: Table row, Column row, missing value, special character ..
    5) 필요 데이터 추출
  - Data evaluation
    1) SQL script execution
    2) SQL result 평가 테이블에 저장
    3) Categorized numerical data according to Data type
  - Posted computed
    1) Time trend, Conditional data ratio summary
    2) Linear regression analysis using Continuous numerical data
    3) 3 SD analysis using Categorical data
    4) DQ score

- DQ Tool :: Meta part

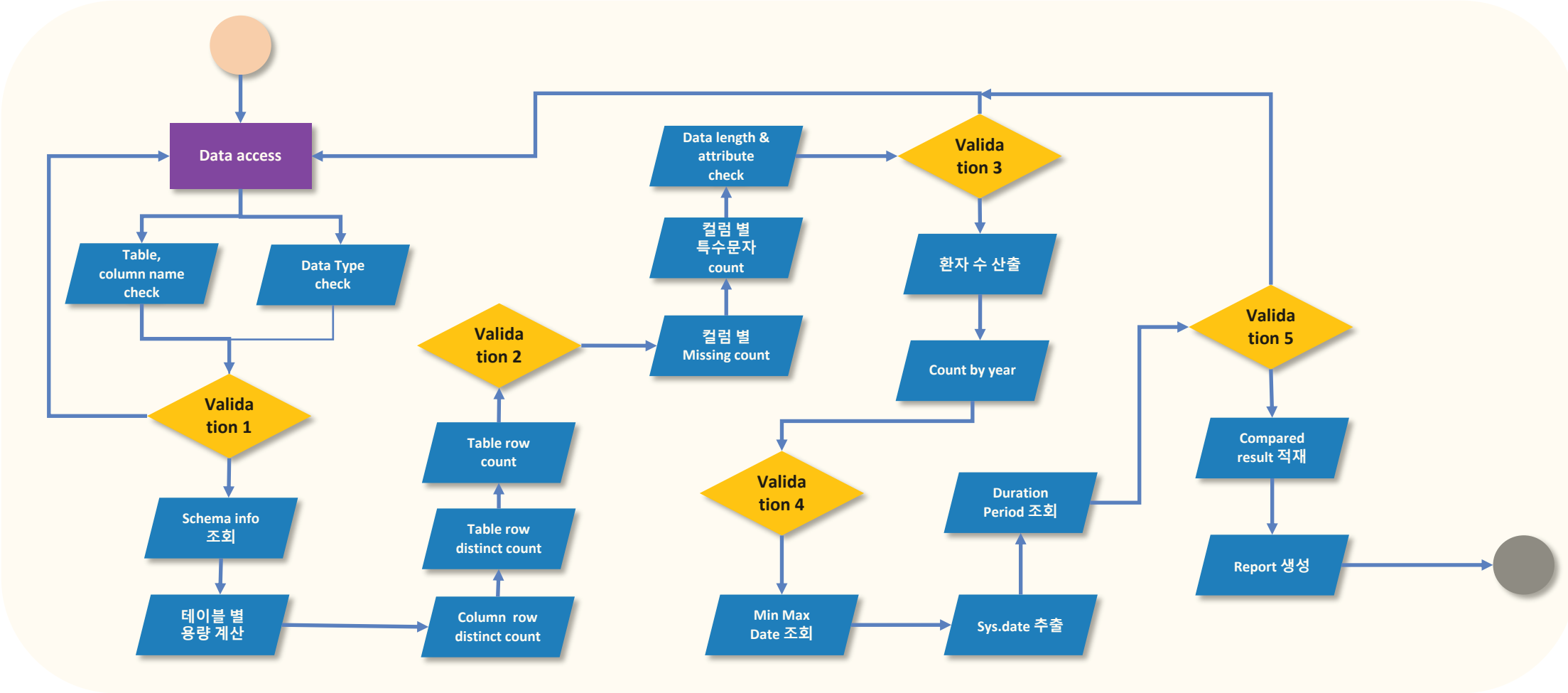
3. system running process

  - Pre computed -> Data evaluation -> Post computed -> Visualized -> Run dash board
    - Visualized
      1) Tree map
      2) Cytoscape (Data base relation..)
      3) Count result table
      4) Regression, Box plot, dispersion, histogram, 3D matlab plot ..
      5) Txt message table
    - Run dash board
      1) Execution shiny
      2) Data quality Score and txt message (Data grade..)
      3) Visualized plot
      4) Txt message

- DQ Tool :: Meta part

  3. system running process

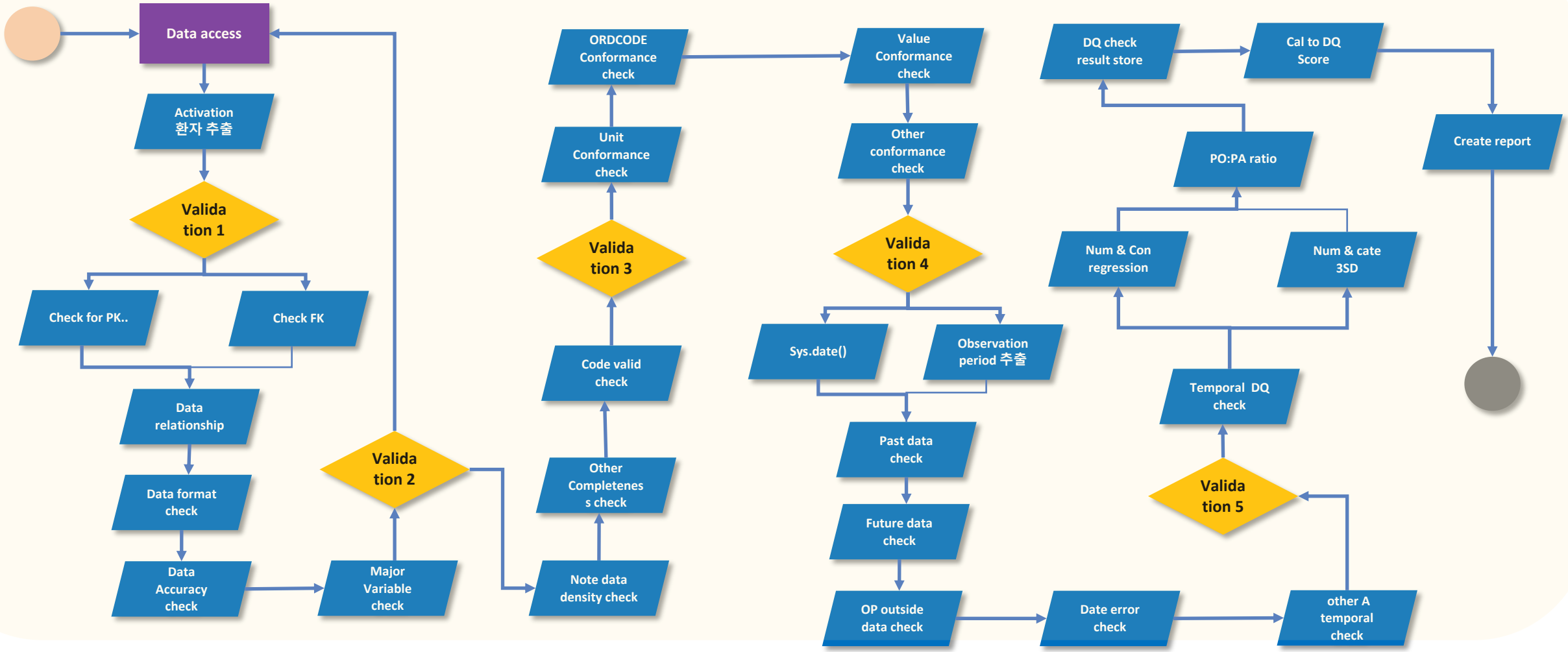
     - Pre computed



- DQ Tool :: Meta part

  3. system running process

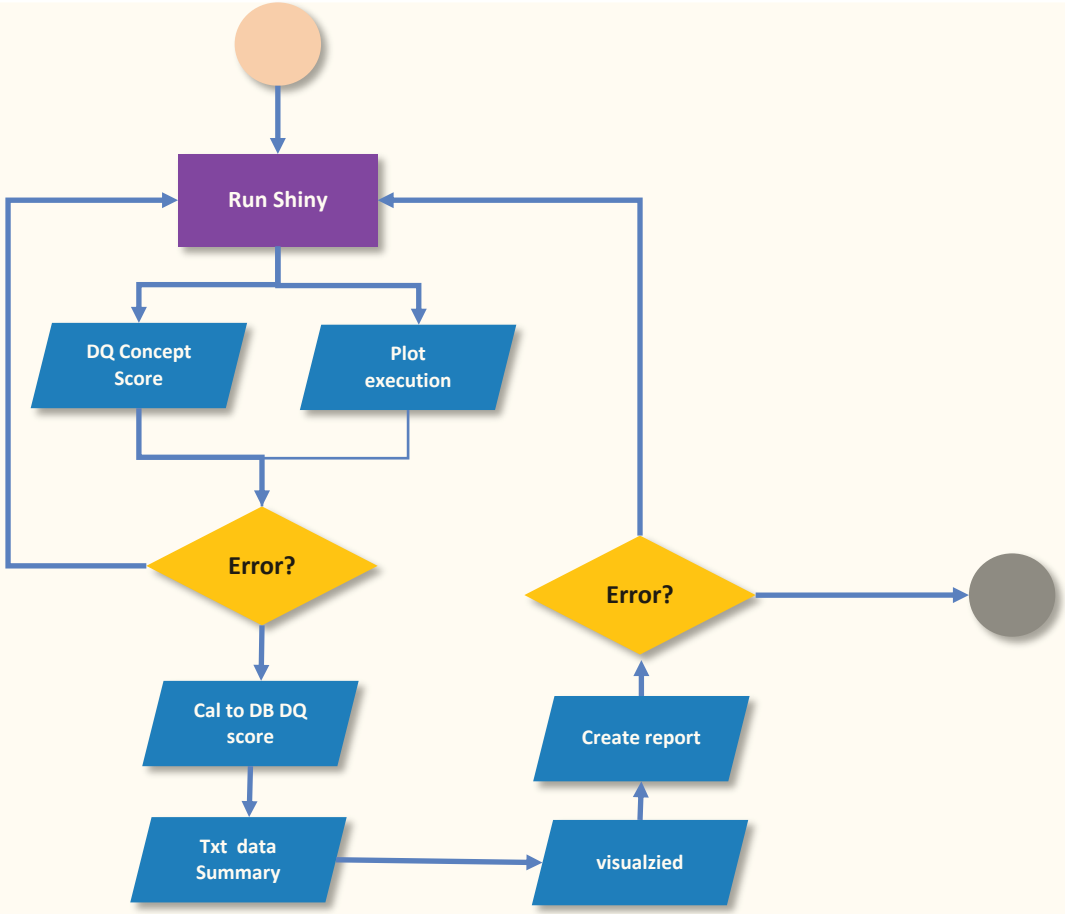
     - Data evaluation



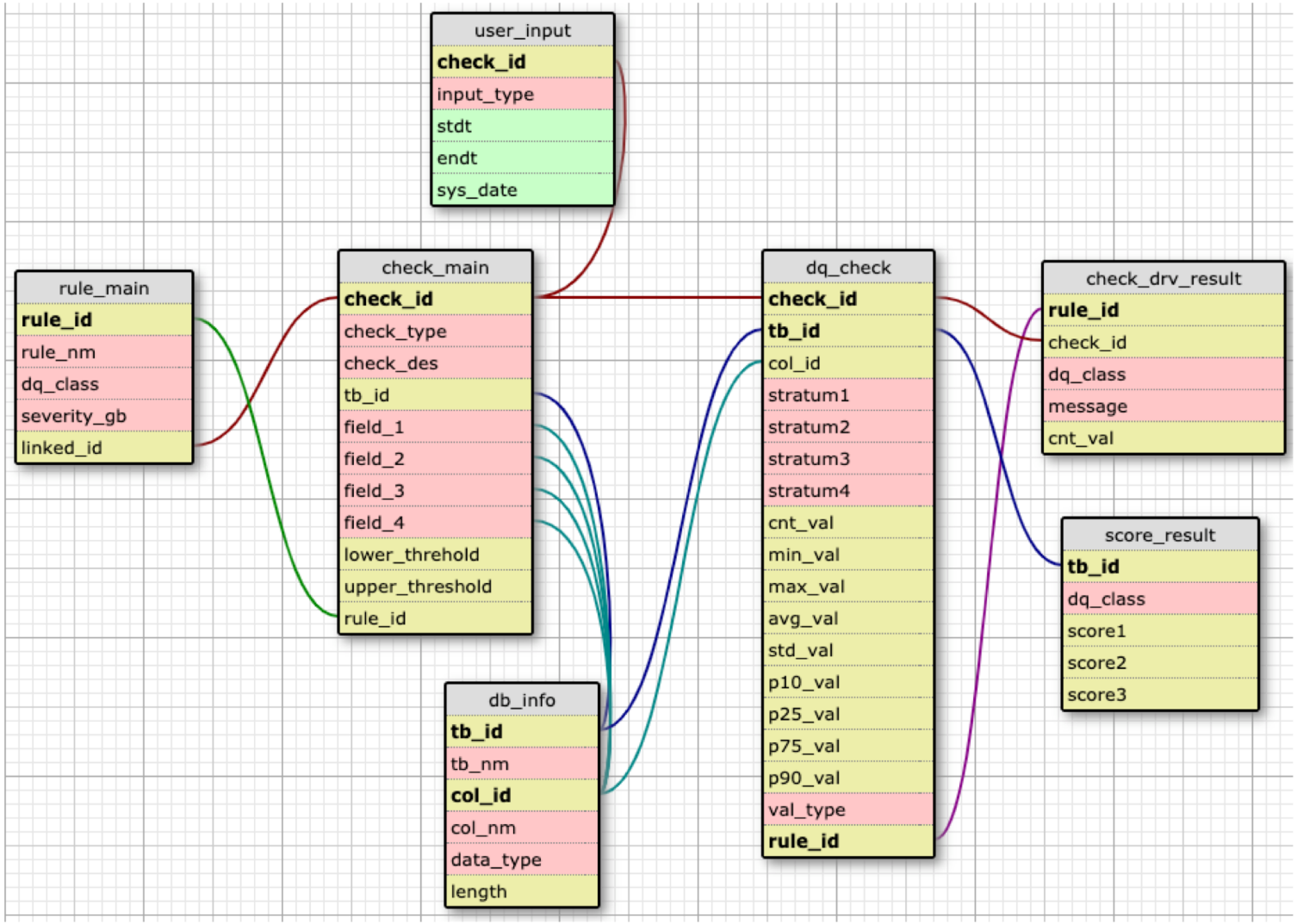
- DQ Tool :: Meta part

  3. system running process

     - Post computed



- DQ Tool :: Meta part
  - DQUEEN Tool's ERD



- DQ Tool :: Meta part

4. Modularity (update on going)

    - Input parameter
        - DB name (Variable input :: sourcename <- DW)
            1) Source data DB name (원본 데이터 디비명)
            2) Meta data DB (메타 데이터 디비명)
            3) CDM DB (CDM 디비 명)
        - Data mapping (Source to Meta, provide csv format)
            1) Meta table name
            2) Source data name
            3) Join type (if join type is null then standard table)
        - Excution level (variable input :: level <- 1 to 3)
            1) Level : 1~ 3

- DQ Tool :: Meta part

4. Sequence Diagram