**Data Profiling**

* Agenda
* Why Data Quality?
* Data Quality common issues and dimensions
* What is Data Quality?
* Introduction to Data Profiling

**Why Data Quality?**

**Heterogeneous Sources of data**

* Different source systems
* Multiple processes & rules
* Manual data entry (Human Errors)
* External data

**Data in real world is dirty**

* Inconsistent data/Duplicates
* Incomplete data
* Ambiguous/noisy data

**Problems caused by bad data**

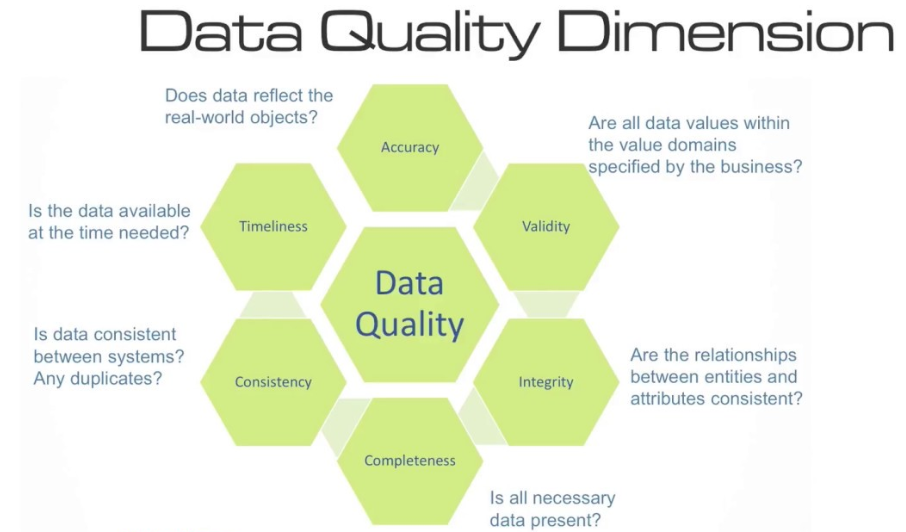
* Loss of credibility in system
* Lost revenue
* Extra costs
* Customer dissatisfaction

**Data Quality common issues and dimensions**

* Completeness
  + Ex: Missing Data
* Integrity
  + Ex : Same city, different zip codes
* Validity
  + Ex: Is this a valid age?
* Inconsistency between systems
  + Ex: 1,2,3 vs. A,B,C
* Accuracy
  + Ex: Smith vs. Smitz

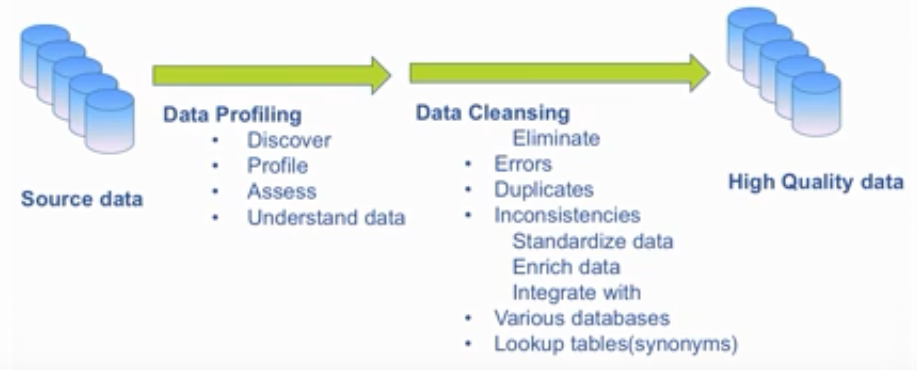
**Data Quality Dimension:**

* Are all the values within the value domains specified by the business?
* Are the relationships between entities and attributes consistent?
* Is all necessary data present?
* Is data consistent between systems? Any duplicates?
* Is the data available at the time needed?
* Does data reflect the real-world objects?



**What is Data Quality?**

-Process of converting source data into high quality data



**I.e. No Quality data=> No Quality decisions**

**Introduction to Data Profiling**

* Enable Users to :
  + Rapidly analyze the data
  + Understand data challenges early in any data intensive project
  + Identify data issues
* Give metrics on data quality including whether the data conforms to particular standards.
* Give visibility as to where certain issues are occurring.
* Increase confidence in your enterprise data

**How Conduct Data Profiling?**

* Simple Statistics
  + Look out for the number of NULL/blank values
  + Check primary key field has distinct values only
  + Check list of possible values for certain fields
* Pattern
  + Check that all values from a field respect a pre-defined format
* Thresholds
  + Define business rules, set thresholds and check anomalies
* Functional Dependency
  + Check relationships between attributes

**Data Profiling**

* **Simple Statistics**
  + Completeness Indicator
    - Identify Column with nulls/blanks
  + Uniqueness Indicator
    - Check unique/distinct values compared to row count
* **Patterns**
  + Patterns indicator
    - Identify different data formats in a column
* **Thresholds**
  + Threshold indicator
    - Define business rules and set thresholds to identify anomalies
      * Ex: Check if all customers are more than 18 year old , threshold is 100%
* **Functional Dependency**
  + Check Relationships between attributes
    - Business dependency between fields
      * Ex : Zip code and city must always be inline

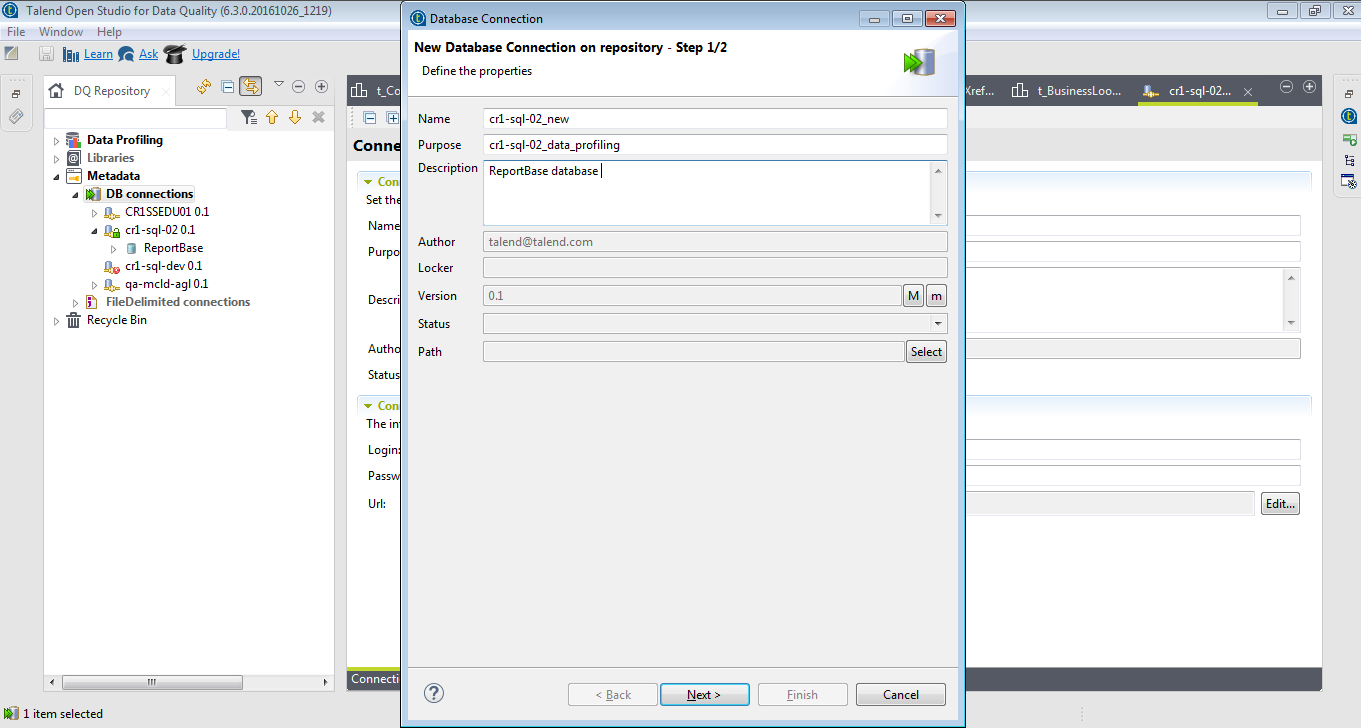
**Data Profiling Demo**

**Tool:** Talend Open Studio for Data Quality

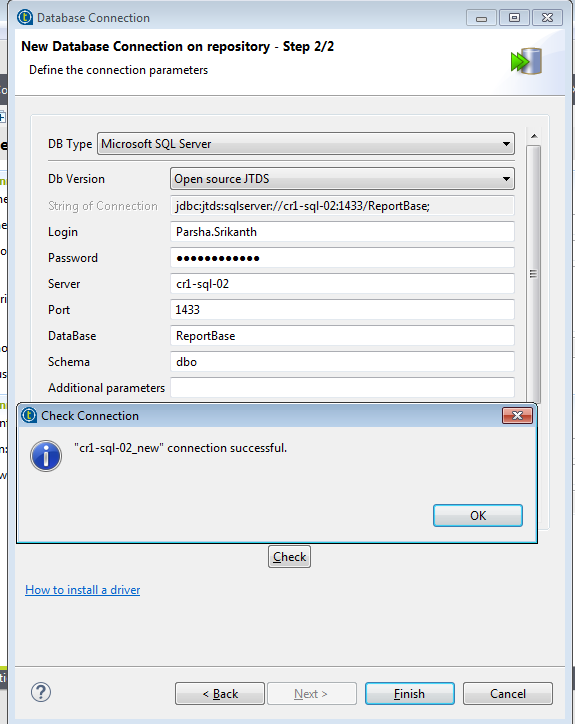
**Data Source:** Microsoft SQL Server

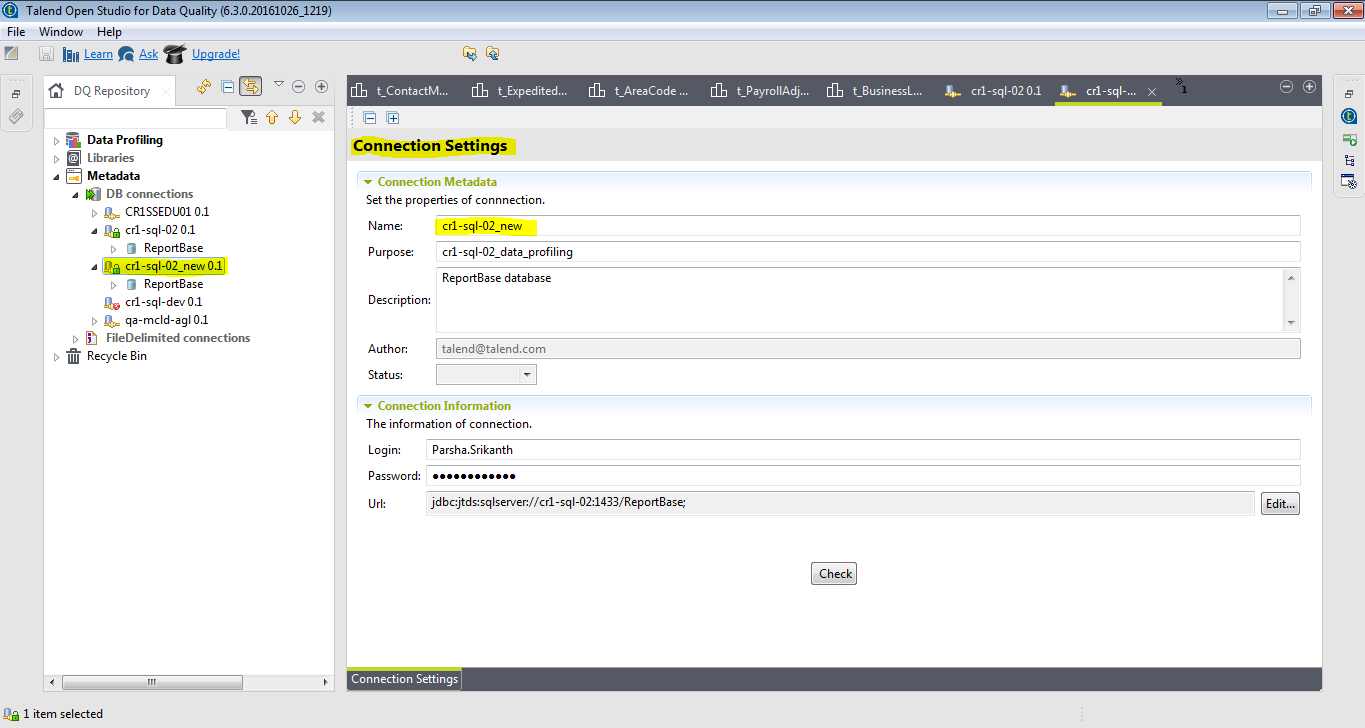
**Step 1:** Connect to the source system (RDBMS: Microsoft SQL server) as below

Provide the metadata in all the fields and click on next button

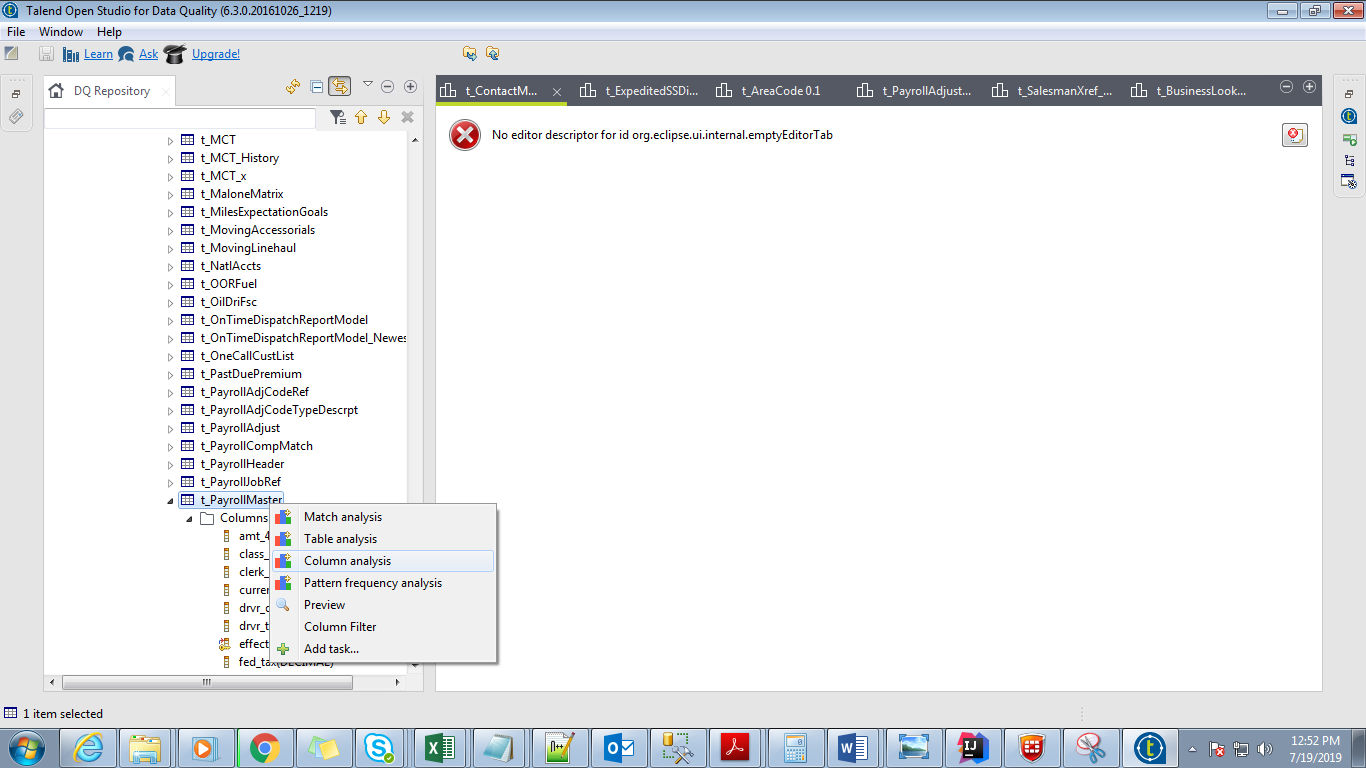


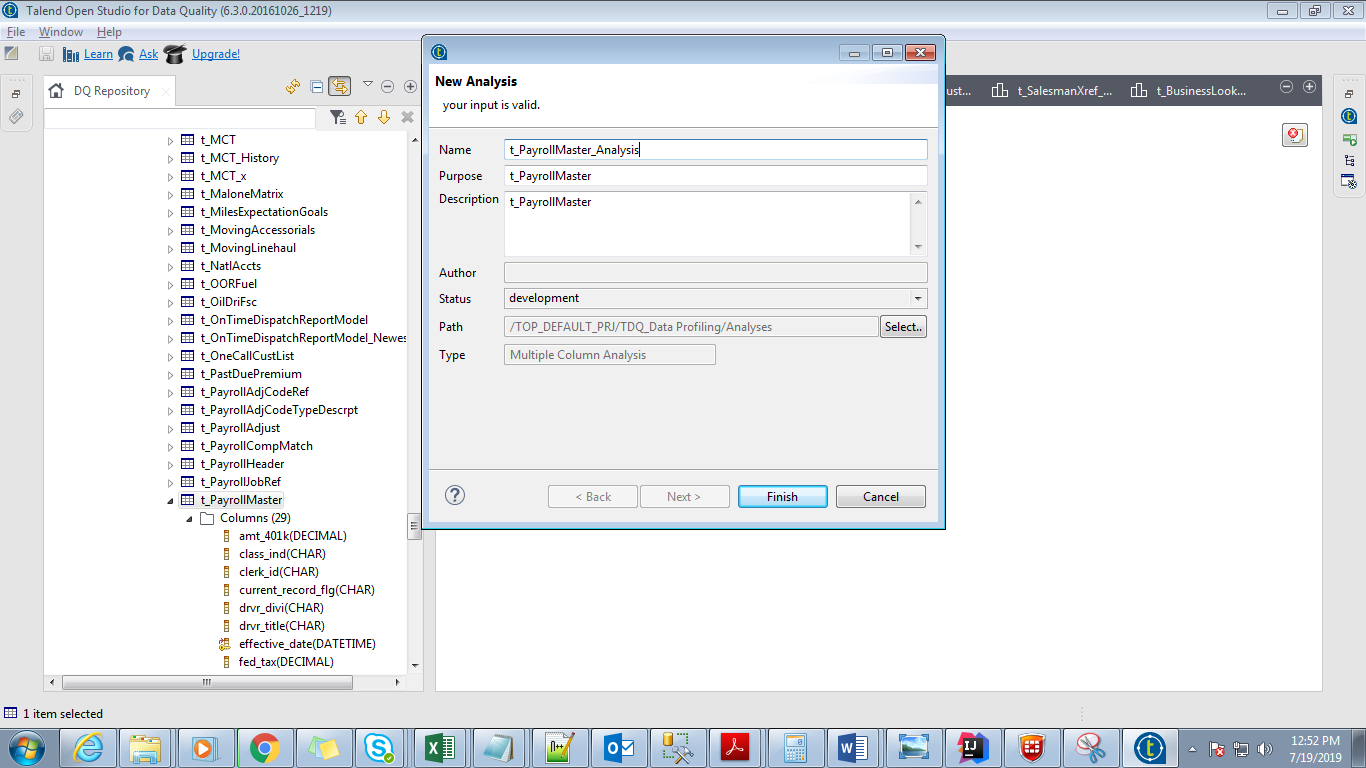
Pass the source connection details as below, verify the connection as below, and then click on finish



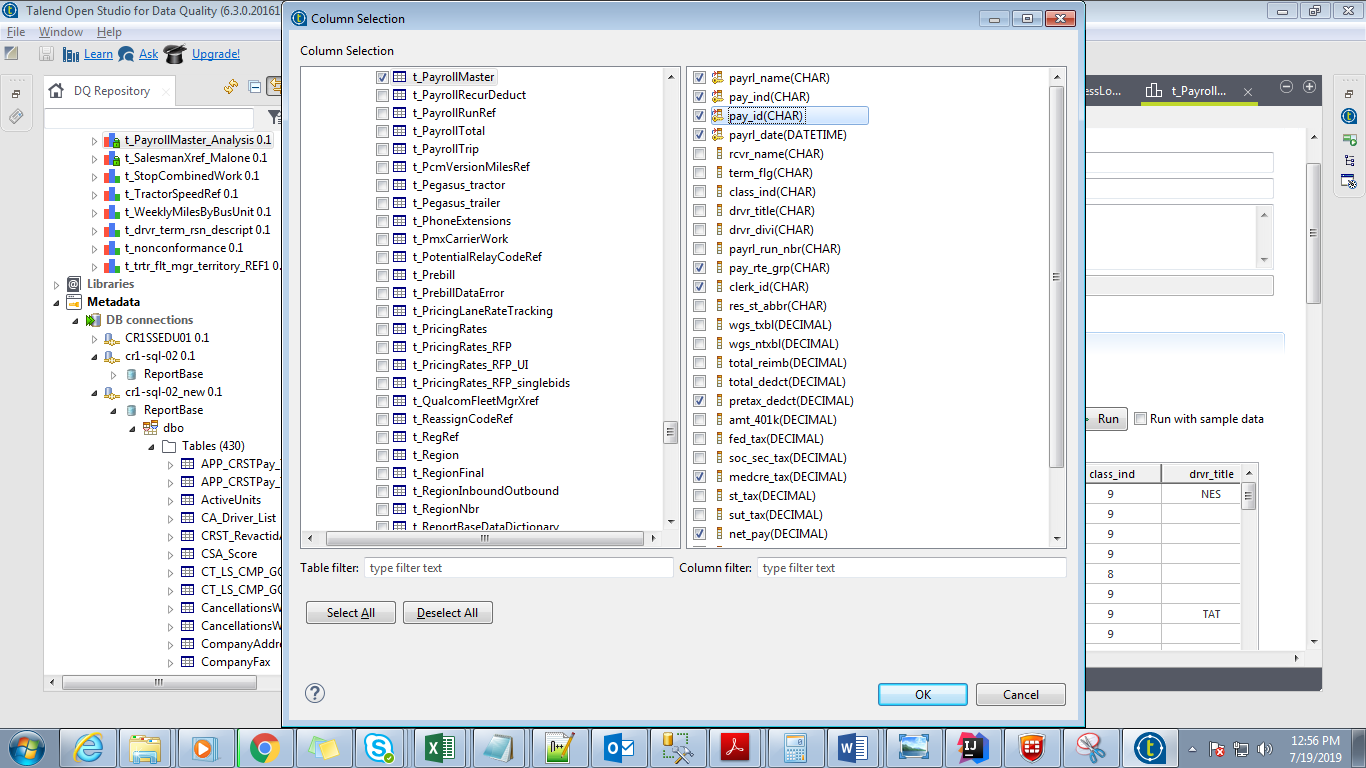


**Step 2:**  Select the specific table and right click and select column analysis, provide metadata in all the fields and click on finish

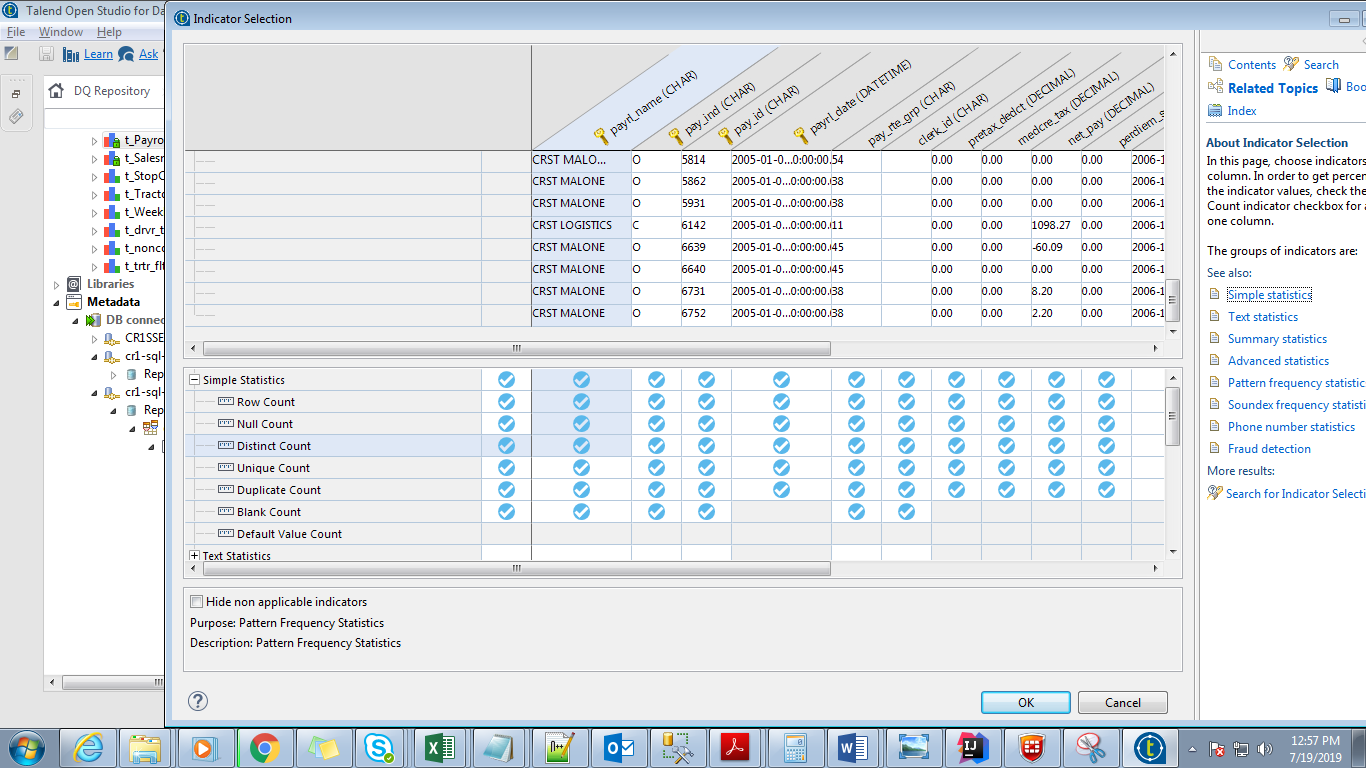


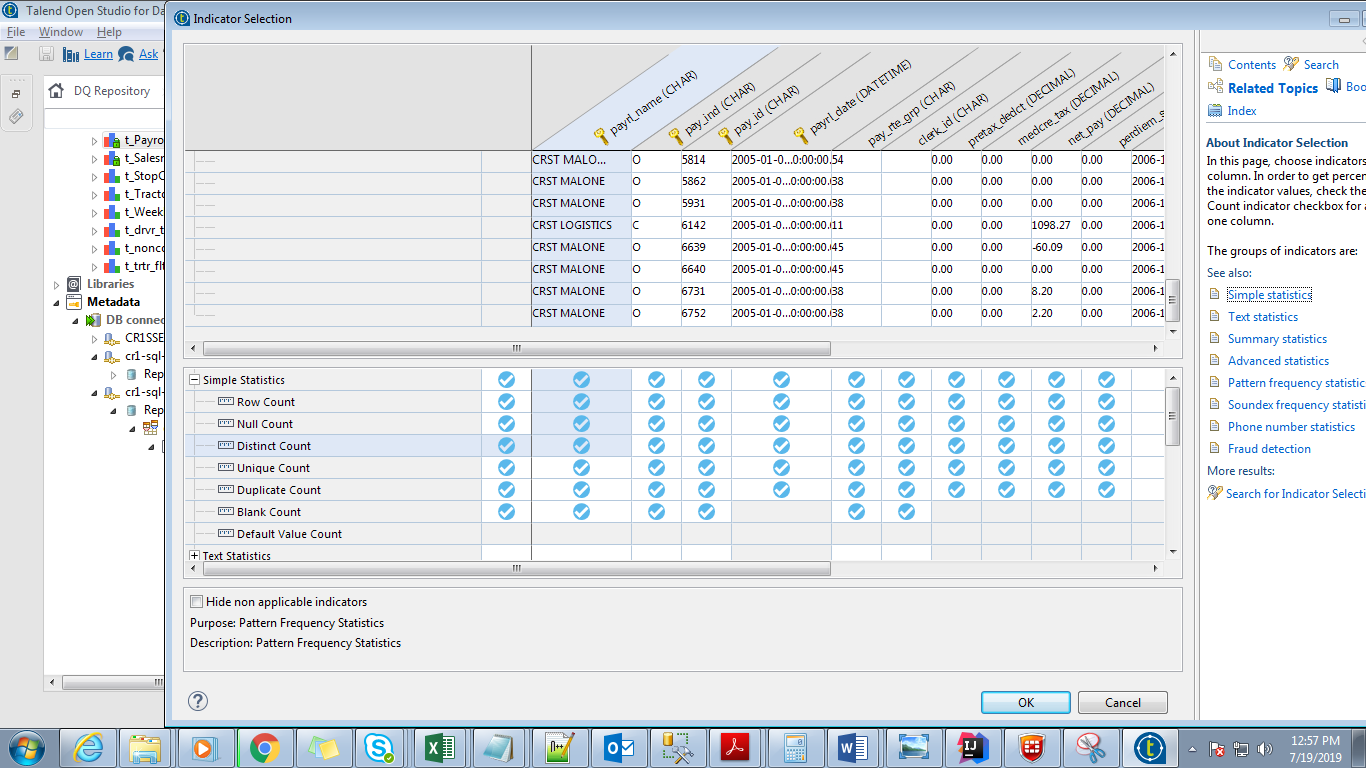


**Step 3:** Select the columns to do the profiling

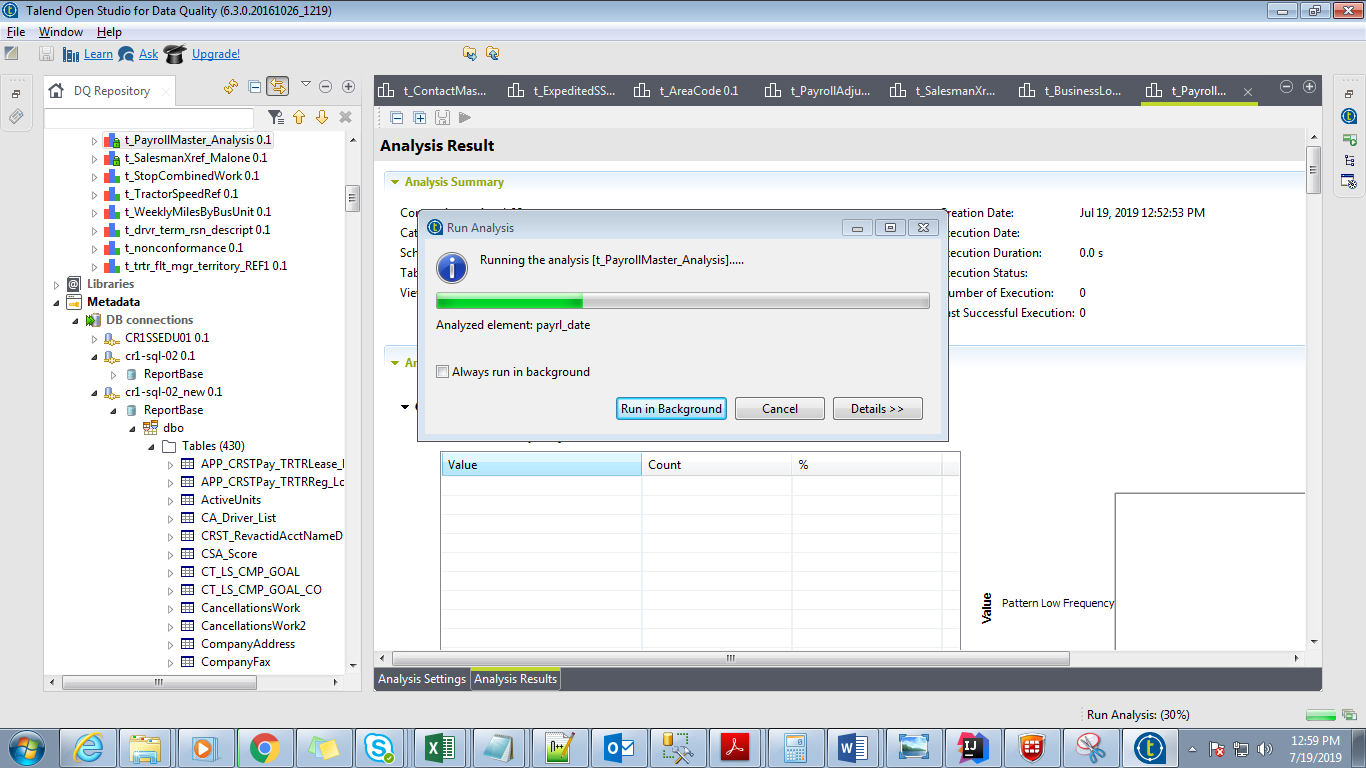


**Step 4:** Select the Indicators

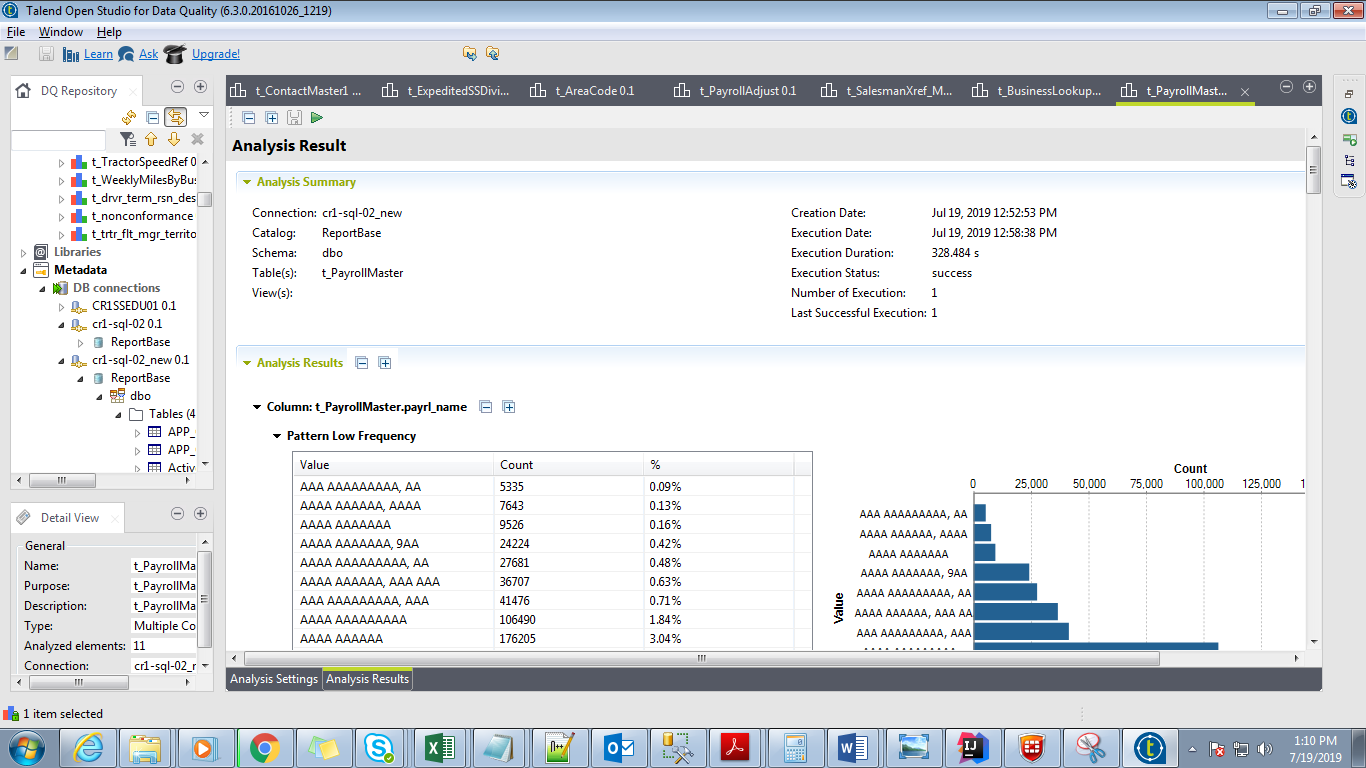




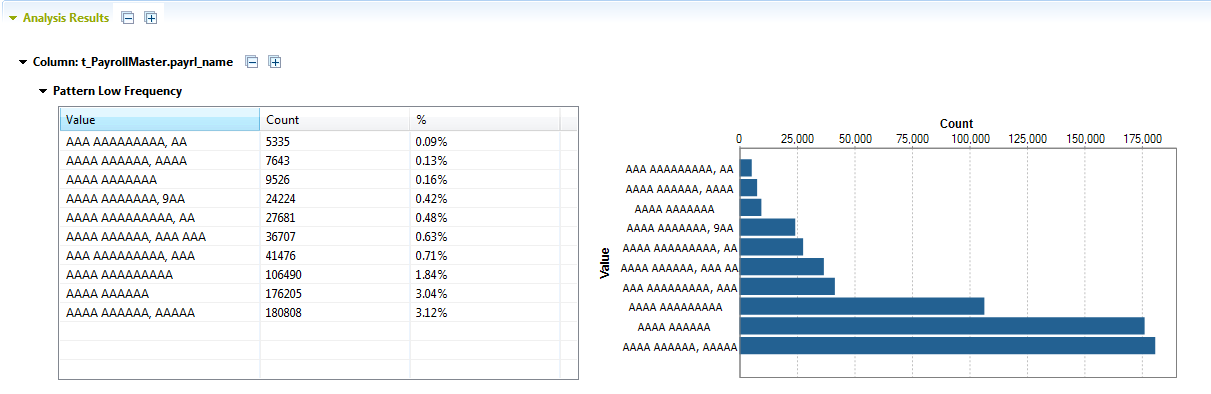
**Step 5:** Click on Run Analysis button



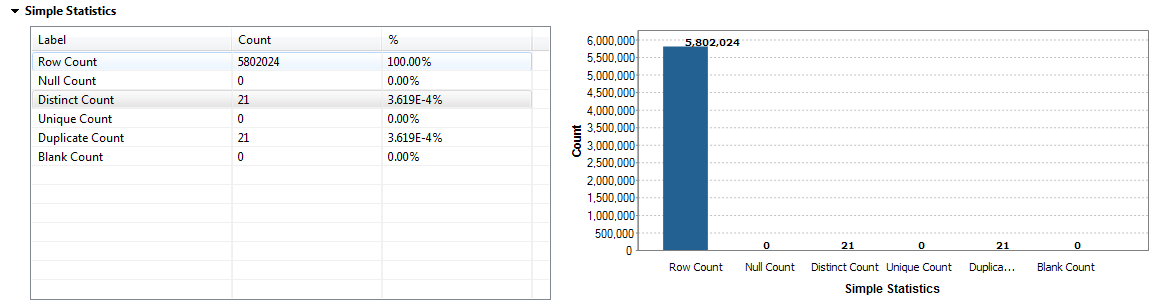
**Step 6: Analysis Result**

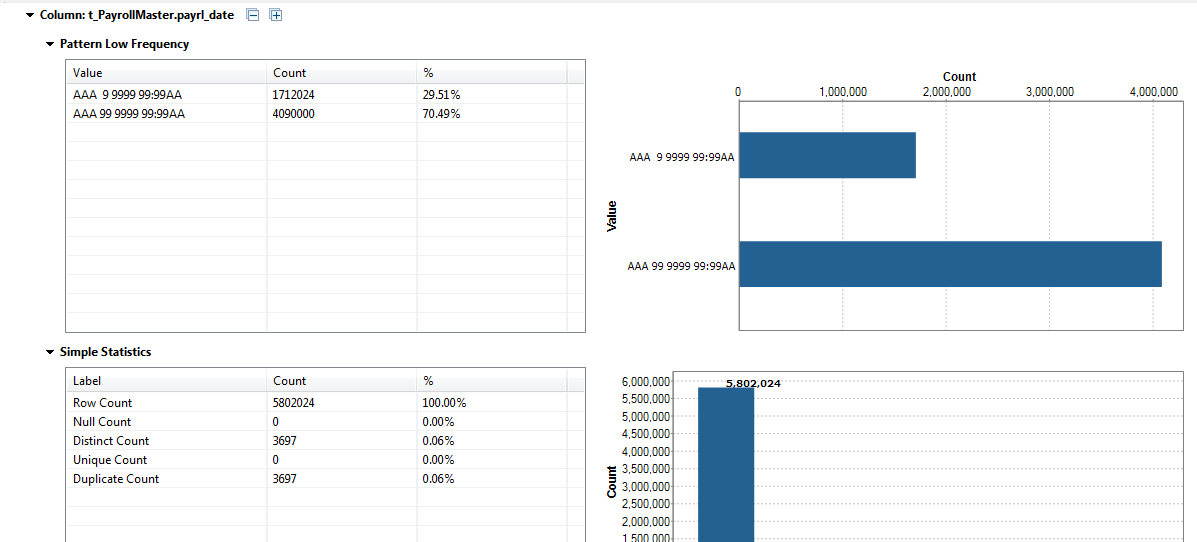


**Pattern Frequency:**



**Simple Statistics**





**List of Tables for Data Profiling**

|  |  |  |  |
| --- | --- | --- | --- |
| **Server Name** | **Database Name** | **Total No.of Tables** | **Status** |
| cr1-sql-02 | ReportBase | 318 | Completed |
| qa-mcld-agl | lme\_qa | 16 | Completed |
| cr1ssedu01 | DBSglep | 2 | Completed |
| DBSjepc | 2 | Completed |
| DBSrrcv | 4 | Completed |