Data Engineer challenge - Results

Examination of the given data

Patents, standards and declarations were the three primary datasets examined. From the description available, it is learnt that all the three datasets are interrelated. Of the learnings the underscores in brief were that the patents are the new inventions claimed by a specific company, the new invention will then be examined by a participatory advisory committee comprised of key stakeholders to arrive at a consensus if the patent can be enlisted or augmented to the existing standard. Upon the accreditation by the key stakeholders a plea is made through the Standard Setting Organisation to make a declaration. Declarations are the key to have the patent being included in the standards. The current datasets pertaining to the a fore mentioned aspects are of the telecommunications industry.

Data model

Entity Relationship Diagram was used to formulate a data model for the given datasets. This diagram was plotted using draw.io

Entity Relationship diagram is a popular method adopted to detail the casual relationships between the datasets. This diagram uses two linking elements from the datasets to establish the relationships. The two linking elements are primary and the foreign key.

One can refer the primary key as the stable identifier. Primary key is one variable identified from the data that is unique to every record in a given table and refers back to one variable and corresponds to one row in a dataset. The variable that is assigned as primary key (PK) does not contain any blank/missing rows specific to that column. Once primary key is identified for one table and the same variable occurs in another table then the primary key is labelled as foreign key (FK) for the other table.

In the ERD plotted, we have three entities that are patents, declarations and standards each of which are depicted by a rectangular box and contain all the headers of the variables from the corresponding tables. These boxes are joined using a line. The edges of the line in this diagram describes the entity occurrences in relation to corresponding table. From ERD, it can therefore be established that one patent may have multiple declarations and one declared patent may lead to multiple standards. This ERD now elegantly represents the relationship between the entities.

DDL and Application for loading data into MySQL DB.

As once after the model has been designed, I have transformed the model into DDL in SQL statements and through which DB is created.

An application was developed using the Spring Batch framework in Java, through which the data which is in CSV format are loaded into MySQL DB. I have used Spring tool Suit 4 develop and deploy the application.

Data analysis

After loading the data there are some insights from the data which were quired, to start of with I have put simple analysis and later focused on the major insights.

- From patents table we could analyse that every year which company has most number of applications, and we could find out that Samsung Electronics Co. Ltd. has applied more number of patents when compared to other companies.
- From patens we could get an insight in which respective year most patents, we could find out that in 2013 highest patents with count of 10 and in 2016 – 6.
- We can fetch the applicants for a given year 2012, the results demonstrate that two companies, Samsung Electronics Co. Ltd. and Apple Inc. were the applicants in the year 2012.

- It is observed that the patent status of many publications has been lapsed.
- On filtering the variable family size of the INPADOC for less than and equal to 2, there were 7 observations.
- On exploring the applicant that exceeds a market coverage of 1.5, Samsung Electronics Co. Ltd. was found to be the only applicant with the market coverage exceeding 1.5.
- The temporal trend of the declarations revealed a drastic dip in the number of declarations for the year 2018 from 2017. However, the highest number of declarations was recorded in the year 2019 as compared to the previous years (2017 - 2019).
- Among the declaring companies, i examined the highest declarations by declaring company per year. Over the years (2011 - 2019), the highest number of declarations by company in an year was Samsung Electronics Co. Ltd. in the year 2019.
- By reducing the redundancies, 46 unique records were identified on merging three data tables based on the identified primary keys.

There is a sample analysis done using Python and could be beneficial to know the insights.