ISSN 2347 - 3983
Volume 8. No. 5, May 2020
International Journal of Emerging Trends in Engineering Research
Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter40852020.pdf
https://doi.org/10.30534/ijeter/2020/40852020

Business Intelligence for Explore Customer in Internet Movie

Abba Suganda Girsang[1] , Arief Handany[2], Christopher Edmond[3], Marcellino[4]
[1]Computer Science Department, BINUS Graduate Program – Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480, agirsang@binus.edu
[2]Computer Science Department, BINUS Graduate Program – Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480, arief.handany@binus.ac.id
[3]Computer Science Department, BINUS Graduate Program – Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480, christopher.edmond@binus.ac.id
[4]Computer Science Department, BINUS Graduate Program – Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480, marcellino003@binus.ac.id

ABSTRACT

IMDb (Internet Movie Database) is an online database website that contains the information for movies, videos and tv programs. The databases contain many demanded valuable data for movie stakeholders in industry. Building a data warehouse is very important for data analysis that can help stakeholders to collect data and transform it into knowledge using Business Intelligence tools. In this paper, we will develop the data warehouse architecture for IMDb using kimball methodologies specifically for making a reporting system to analyze actor data like popularity and movies statistics. The dashboard results are expected to be valuable assets to be used by movie stakeholders and future data analysis.

Key words : Business Intelligence, Data Warehouse, Dashboard, ETL, Online Database

1. INTRODUCTION

IMDb (Internet Movie Database) is an online database website that contains the information for movies, videos, and TV programs [1]. While IMDb website is managed by IMDb.com, Inc. staffs however the data is provided by the volunteer contributors and then it will be verified by the editors which are part of IMDb.com staff. People on the internet can visit the IMDb website to find the information about their favorite movie, they can also comment and give a rating for the movie they watched. Not only details about the movie, it also contains the crews and actors that play roles for the film, the movie's language, the movie's genre, and any other detailed information regarding the movie. All of this data is gathered and donated by IMDb's volunteer contributors.

As one of the most popular sources of information for movies, IMDb collected many valuable information such as audience preferences and even actors' popularity. Movie industry has a huge demand for this kind of information especially to create a new movie is a huge investment and you need to know your audience first, additionally you also need to know the best actors to play a specific role for your new movie. And so building a data warehouse is a natural progression from the previous iteration of the database, using data warehouse analysts can rapidly extract data and analyze it using a BI (Business Intelligence) tools which can reveal critical knowledge in business. BI systems are an opportunity for organizations to quickly and effectively use information and transform it into useful knowledge that enables them to meet business objectives [2]. The meaning of BI in real-time depends on understanding and agreeing with things with real-time for business [3]. Data warehouse is the most important component for the decision making process that can ensure the success in creating business intelligence [4]. Data inside the data warehouse can be utilized for applications such as dashboard to perform data analysis [5]. One of the most popular methods in building data warehouses is kimball methodology and one that will be used in this study [6].

In this paper we will develop a data warehouse for IMDb and IMDb reporting using Tableau BI tools [7] based on several considerations including fast data processing, beautiful visualization, and compatibility with kimball methodology. Extract, Transform, and Load (ETL) or data integration process will be done using Pentaho [8]. Data warehouses have been well developed in other kinds of industries such as higher education [5], climate analysis [9], supply chain [10], and many other examples [11, 12]. This paper is based on the kimball lifecycle until deploying the BI [6]. The end result will be reports that can be used by the movie stakeholders that specifically gathered from IMDb website.

2. LITERATURE REVIEW

This section presents the in-depth analysis for literature that serves as a basis for our methodologies in chapter 3. We

mainly used kimball methodology [6, 13] as our go to methodology when designing our data warehouse and our BI, these methodologies will be explained in multiple sections.

2.1 Database Definition

There are many definitions for databases but a database is basically a collection of data that fulfill the requirement of business, each data may contain logical interconnection [14]. [15] further elaborate that a database is a collection of data that is stored and can be integrated and centrally managed and controlled.
Based on the definition above, it can be concluded that the database is a logical group of interconnected data and stored to be processed into useful information for the organization. The process of managing and controlling also further differentiate between data and databases. Databases is also a method for organizing sets of data.

2.2 ETL (Extract, Transform, and Load) Tools

According to [6], "ETL is used to migrate data from one database to another, to form data marts and data warehouses and also to convert databases from one format or type to another."According to [16] there are three process of ETL:

- Extraction: Extraction process in the first phase, the data will be retrieved from internal and external sources. Logical differences will occur at the beginning of extraction, where different data related to past data entered into the data warehouse is empty, and the extraction will be done to update the contents of a data warehouse that will be combined with the new data that are available from time to time. Selected data will be taken depending on the design of the data warehouse issued, which will be tailored to the needs of business intelligence analysis and decision support systems.
- Transformation: The purpose of cleansing and transforming processes are aimed to enhance data quality. In order to improve the accuracy and quality of all the data collected from multiple sources by fixing current inconsistencies.
- Loading: After extracted and converted, the data will be loaded into the data warehouse tables so the data can be used for analysis and creating business intelligence reports. Also clearly spelled out by [16] Applications acquisition, which is also called extract, transform and load (ETL) or data integration tools, is an application that allows a set of data to be extracted, transformed and prepared to be loaded to the data warehouse. Business intelligence or decision support systems are front end applications that allow experts to evaluate and interpret the visualization based on the outcomes.

To do ETL process we need data integration tools like Pentaho [8], we choose Pentaho because it is easy to understand user interface and fast batch processing which is very compatible for building a data warehouse from raw data. Based on our analysis Pentaho is also compatible with kimball methodology for ETL process. [13]

2.3 Data Warehouse

According to [17], "A warehouse is a subject-oriented, integrated, time variant and non-volatile collection of data in support of management's decision making process" which based on [18] can be breakdown into four aspects:

- Subject-oriented: Data that is separated by function in the business process.
- Integrated: Data obtained from multiple sources and can be integrated to be analyzed together.
- Time-variant: Data that contains historical data for each transaction inside.
- Non-volatile: Data that is read-only which means it won't change the real data.

In summary, a data warehouse is a centralized database that contains business data that can be used for analysis. The main challenge in data warehouse is to gather, collate, and build large amounts of data from many systems for collection based in business intelligence, the process of loading data in a data warehouse done after performing the Extract, Transform, and Load (ETL) processes [19].

2.4 Granularity

According to [17],"Granularity Refers to details of units of the data in the data warehouse." The deeper the data is, the lower the granularity point. The less detailed data, the higher the granularity level. Granularity declaration is an important aspect of the Kimball empirical data warehouse modeling framework for data warehousing [20].

2.5 Star Schema

According to [14], "Star schema is a logical structure that has a fact table that contains data in the central fact table surrounded by dimension-table containing the reference data or evidence can usually be denormalized." According to [16], "based on the representation of multidimensional star schema that contains two types of data tables are dimension tables and fact tables." From the definition it can be concluded that star schema is the logical structure of the data that enables two types of tables are dimension tables (dimension tables) and fact tables (fact tables). Dimension tables contain data that is compatible with the needs of business and the facts surrounding the table to obtain the information.

According to [16], the star schema has two kinds of tables: fact tables (fact table) and table dimensions (dimension table). Fact table(the fact table) is a table that generally

contains something that can be measured(measure) and historical, and a collection of primary key, foreign key contained in each dimension table. Usually refers to the fact table transactions and contains two types of data:

- Link to the dimension tables, which is required as a reference of the reference information contained in each fact table.
- The number values of attributes that define a transaction that includes and represents the real target of OLAP (Online Analytical Processing) analysis.

Therefore, the fact tables contain derived data and connect one or more dimension tables. Table dimensions(dimension table) are a table that contains categories with summary details of which can be in the form of reports. In general, it can be called a dimension associated with surrounding entities in the organization process.

2.6 Dashboard

Dashboard is a direct descendant of the old EIS and DSS systems, by improving the functional and performance [21]. Because they are connected with strong data systems and utilize Key Performance Indicators [22]. three types of dashboard such as :

- Tactical Dashboard: Measure and describe the short-term productivity and effectiveness of the company's performance. The result is the dashboard displays that can describe the company in making a smart strategy.
- Operational Dashboard: Measuring the effectiveness of business functions running on a team or business unit level. Dashboards of this type can potentially be used by a team manager.
- Strategic Dashboard: Built to carry out the policies of the organization. Dashboard displays data describing business strategy and corporate goals.

2.7 Business Intelligence

BI (Business intelligence) is allowing users to transform data into meaningful data that can support decision making processes in the company while data warehouse is the most important component in creating BI [18]. [6] further elaborate the varieties of BI applications, BI can be a simple query tool or a complex data modeling tool however all of them help to present data in meaningful ways regardless of the input method of data or visualization produced from the data. Based on the illustration above, we can conclude that a data warehouse is the source of data or the kitchen while ETL processes the data like a cook and also moves the data like a waiter moving food to the customer. BI is the dining table where data is ready to be served [23]. We will use Tableau as our choice of BI because of the compatibility with kimball methodologies [24].

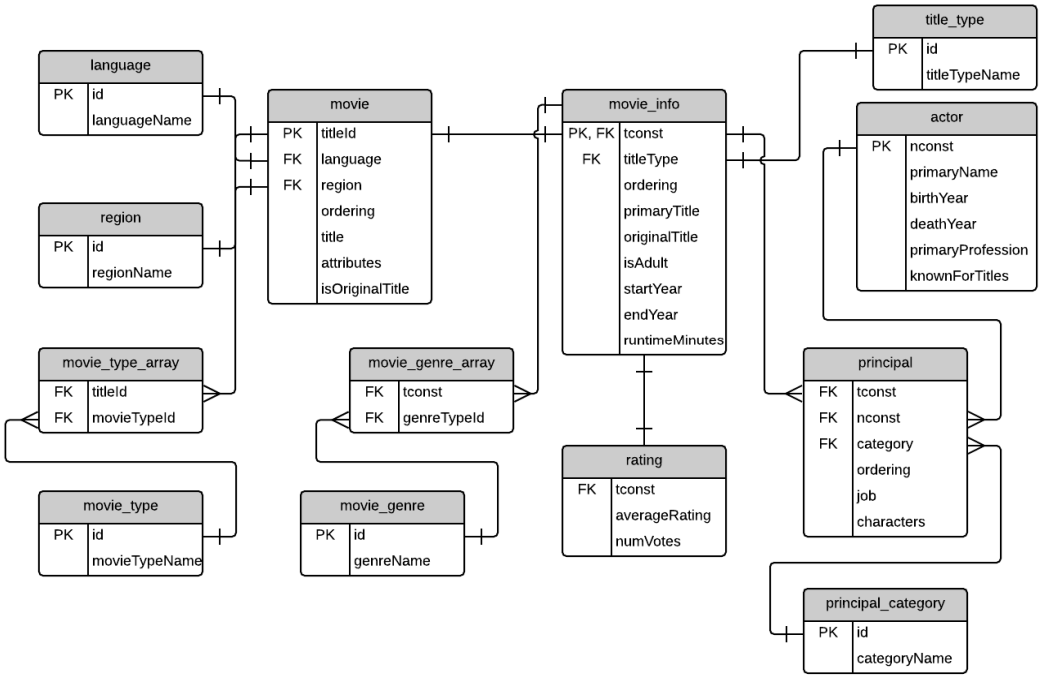

Figure 1: IMDb Entity Relational Diagram

3. METHODOLOGY

Figure 1 shows the IMDb website's database, we can see that from one perspective the most prominent transaction table is movie and principal. Movies represent the data entry for movies, videos, and short films for IMDb websites while principals represent the relationship between the movie and the actor that plays that movie. Using kimball methodologies we want to create a data warehouse for actor dashboard and in kimball methodologies there are four steps of basic dimensional design process [20] which are:

Select the business process: In this step we need to identify which business process that will be incorporated into the data warehouse, business process usually translate into fact and most fact usually focus on a single business process.

Usually in this step we actually choose the data mart that we desired [6] but because we only have one database, so we only have one database to represent our business process.

- Declare the grain: In this step we will need to declare the granularity of data which we actually use per snapshots of principal (Year time dimension) to make an actor dashboard. Principal represents the history of each actor playing a certain movie, and the grain we choose is the rating of the actor.
- Identify the dimensions: After deciding the grain, we can find the dimension that we need based on the grain in this case the record. As we can see in the above figure, principal contains the actor id (nconst) which means Actor is a dimension. Principal also contains tconst or movie is which means we can collate movie and movie_info table together (because we only want to know about the actor's performance), in the movie there are connected with Genre, Language, and Region table which can be transformed into dimensions. Year in movies can also be used as a time dimension.
- Identify the facts: As pointed out above we want to find the rating/popularity of the actor because we only want to focus on that, we will use Principal as the main fact table.

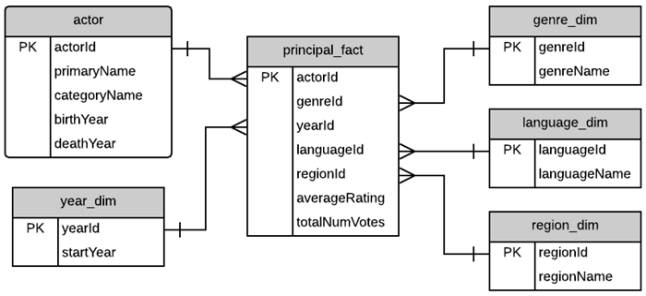

Figure 2: IMDb Star Schema

Figure 2 shows the star schema resulting from the following steps. After creating all dimension tables and fact tables, the data is processed using Pentaho Data Integration tool. The ETL process is shown on Figure 3 – 9.

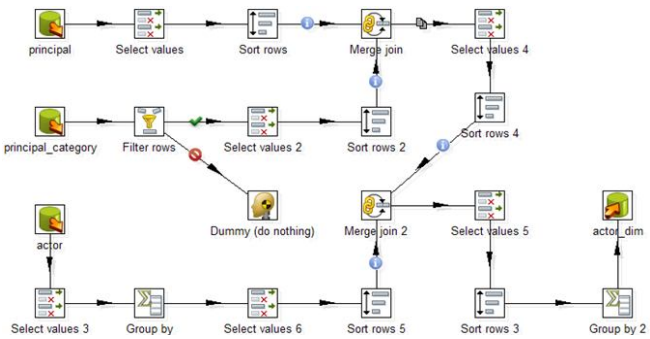

Figure 3: Actor Dimension ETL

In figure 3, we joined data from principal table with principal_category table using filter, to show category only for actress and actor, then the result must be join with table actor to get the birth year and death year.

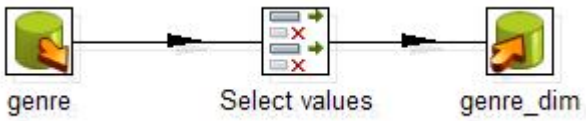

Figure 4: Genre Dimension ETL

In figure 4, we create genre dimension from the genre table from the database by selecting only the genre_id and genre_name.

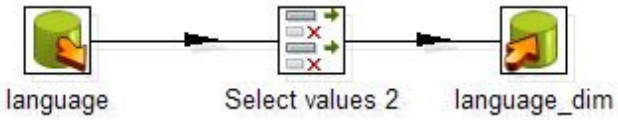

Figure 5: Language Dimension ETL

In figure 5, we create genre dimension from the language table from the database by selecting only the language_id and language_name.

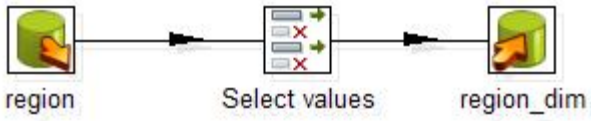

Figure 6: Region Dimension ETL

In figure 6, we create genre dimension from the region table from the database by selecting only the region_id and region_name.

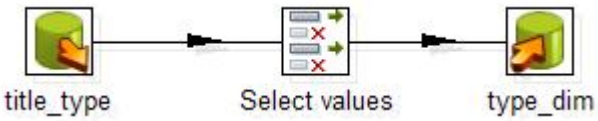

Figure 7: Type Dimension ETL

In figure 7, we create genre dimension from the title_type table from the database by selecting only the title_type_id and title_type_name.

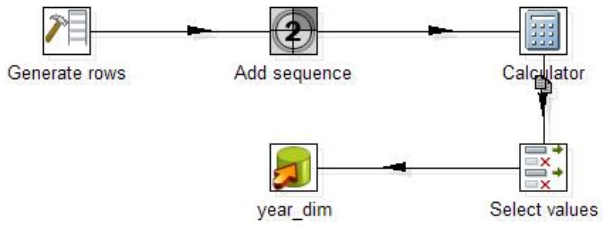

Figure 8: Year Dimension ETL

In figure 8, we create a year dimension, using the generate rows function in pentaho, starting from year 1890 and use a sequence generator to generate 1 value and add it to the year using a calculator.

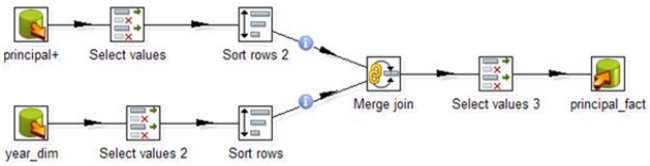

Figure 9: Principal Fact ETL

In figure 9, we create the fact table by joining the principal table with the year dimension, and inside the principal table input, we aggregate numvotes field into total numvotes, and average of rating from the rating field. and load the result into the principal fact.

4. RESULT

After ETL is done, we can proceed to use our data warehouse to build a new dashboard, in this paper we will be using Tableau as our BI tools. After connecting to our data warehouse, we can connect them into a star schema as shown in figure 10.

After the star schema is created, we can continue to make the dashboard using Tableau as shown in figure 11 through figure 14.

As shown in figure 11, we can analyze the distribution of number votes of movies and average between years for actors that were born within 1900 until 1918, as we can see 1901 is where the most votes are but across the year the average movie score is relatively the same.

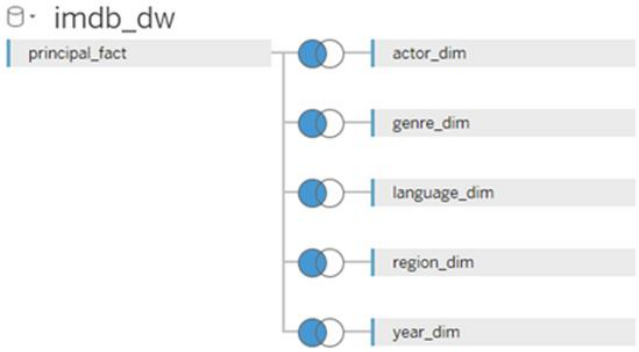

Figure 10: Data Warehouse Connection with Tableau

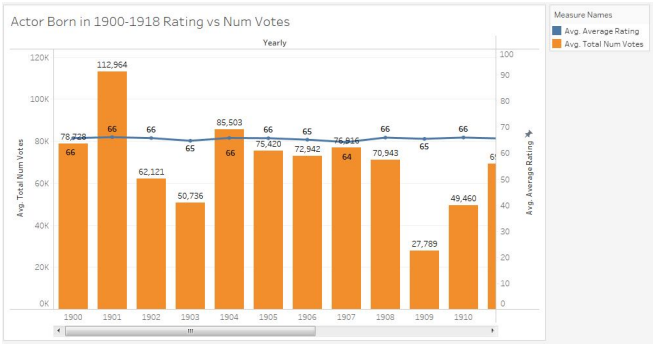

Figure 11: Dashboard Actor Born in 1900-1918 Rating vs Num Votes

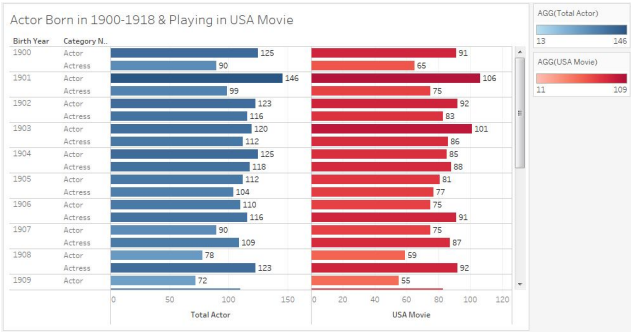

Figure 12: Actor Born in 1900-1918 & Playing in USA Movie

As shown in figure 12 we can analyze the number of actors and actresses from 1900 until 1918, we can also see how many of them are playing USA movies (movie with US region) compared to the total number of actors and actresses. Figure 13 shows the dashboard and their filmography including their average score for each of their movies.

Figure 14 shows filmography detail from figure 13, we can see the average score of each genre and how many votes that represent the ratings. This dashboard is very useful to know the history of actors/actresses and what movie genre they are good at. Complete dashboard can be seen in figure 15.

Figure 13: Actors/Actresses Filmography

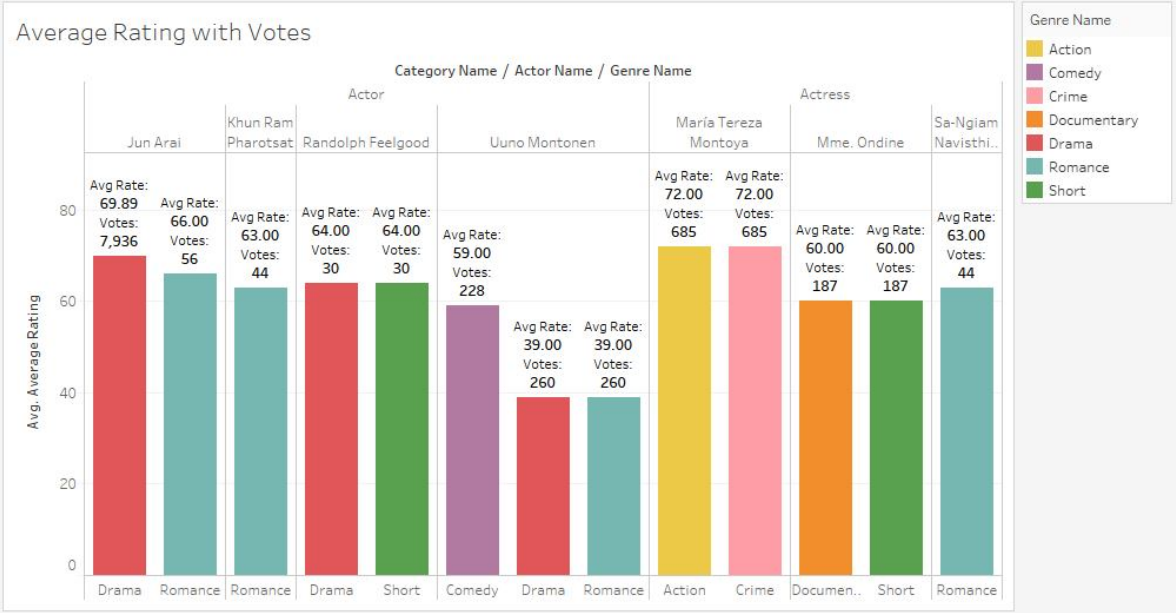

Figure 14: Actors/Actresses Filmography Detail with Average Rating & Votes

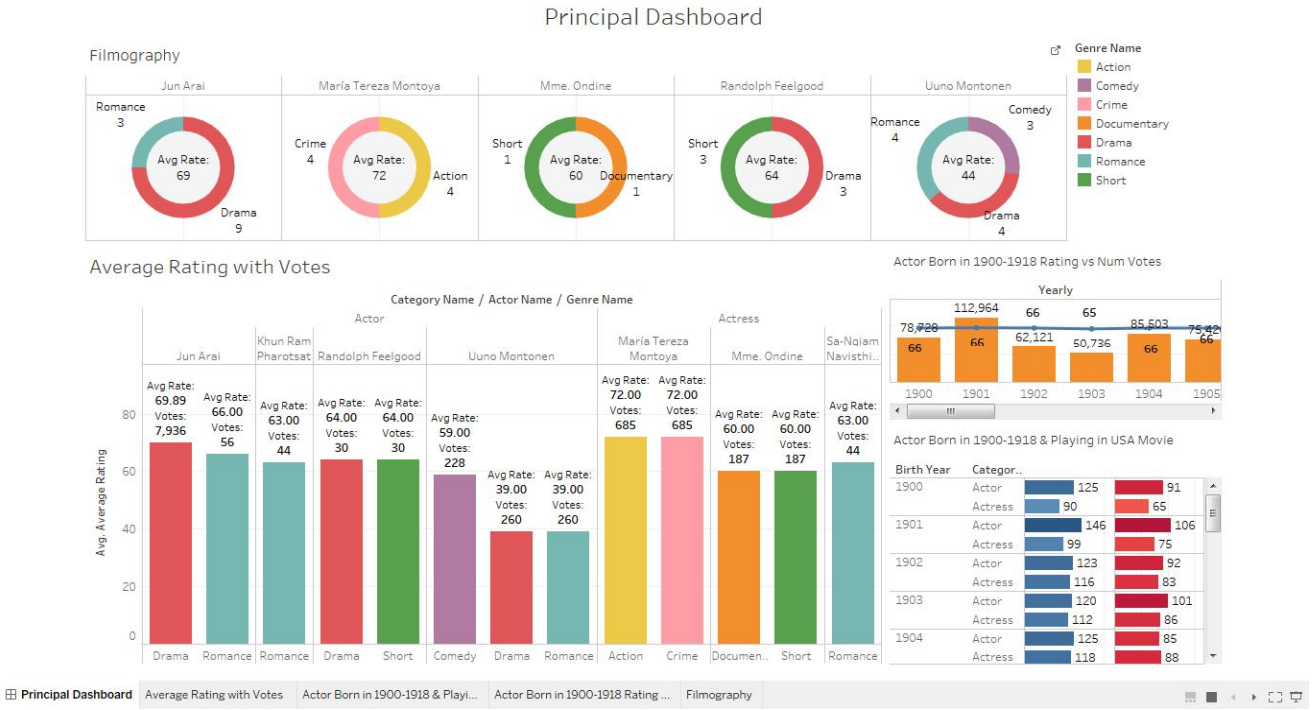

Figure 15: Principal Dashboard

5. CONCLUSION

Data warehouse is an important aspect in leveraging not only a business transaction but also a database-type data entry, by utilizing data warehouse methodologies we can rapidly analyze data and create a reporting dashboard to satisfy the requirement of stakeholders. ETL is also proven as a best method for converting databases into a data warehouse easily without the need to interfere with the production database. Currently we only use Pentaho as ETL tools and Tableau as reporting tools, exploring other tools is also important to gauge the effectiveness of data warehouse methodologies. For future development we can also implement these techniques in big data technology, with a different volume of data, implementing a data warehouse may have it's own challenge and reward but exploring it is a novelty.

REFERENCES

1. **IMDb: Ratings, Reviews, and Where to Watch the Best Movies & TV Shows**, IMDb, available at https://www.imdb.com/.

2. O. Małgorzata, **Business Intelligence as a Future Analysis And Interpretation Of Data In Real Time**, International Journal of Advanced Trends in Computer Science and Engineering, pp. 121–126, Jan. 2019, doi:10.30534/ijatcse/2019/2381.12019.
3. D. W. S. Kusuma, **Business Intelligence Infrastructure of Medical Record Data History System to help Doctorin differencing rare and dangerous disease in patient**, International Journal of Advanced Trends in Computer Science and Engineering, vol. 9, no. 1, pp. 664–672, Feb. 2020.
4. M. Golfarelli and S. Rizzi, **From Star Schemas to Big Data: 20+ Years of Data Warehouse Research**, in Studies in Big Data, Springer International Publishing, 2017, pp. 93–107.
5. L. W. Santoso and Yulia, **Data Warehouse with Big Data Technology for Higher Education**, Procedia Computer Science, vol. 124, pp. 93–99, 2017.
6. R. Kimball, Ed., ***The data warehouse lifecycle toolkit***, 2nd ed. Indianapolis, IN: Wiley Pub, 2008.
7. R. Wesley, M. Eldridge, and P. T. Terlecki, **An analytic data engine for visualization in tableau**, in Proceedings of the 2011 international conference on Management of data - SIGMOD '11, 2011, doi: 10.1145/1989323.1989449.
8. D. Stooder, **TDWI Best Practices: Improving Data Preparation for Business Analytics**, Hitachi Vantara, available at https://www.hitachivantara.com/en-us/pdf/analyst-content/improving-data-preparation-for-business-analytics-tdwi-best-practices-report.pdf
9. Doreswamy, I. Gad, and B. R. Manjunatha, **Hybrid data warehouse model for climate big data analysis**, in 2017 International Conference on Circuit ,Power and Computing Technologies (ICCPCT), 2017, doi: 10.1109/iccpct.2017.8074229.
10. S. Biswas and J. Sen, **A Proposed Architecture for Big Data Driven Supply Chain Analytics**, SSRN Electronic Journal, 2016, doi: 10.2139/ssrn.2795906.
11. A. Sebaa, F. Chikh, A. Nouicer, and A. Tari, **Medical Big Data Warehouse: Architecture and System Design, a Case Study: Improving Healthcare Resources Distribution**, Journal of Medical Systems, vol. 42, no. 4, Feb. 2018, doi: 10.1007/s10916-018-0894-9.
12. J. Campos, P. Sharma, U. G. Gabiria, E. Jantunen, and D. Baglee, **A Big Data Analytical Architecture for the Asset Management**, Procedia CIRP, vol. 64, pp. 369–374, 2017, doi: 10.1016/j.procir.2017.03.019.
13. R. Kimball and J. Caserta, ***The data warehouse ETL toolkit: practical techniques for extracting, cleaning, conforming, and delivering data***. Indianapolis, IN: Wiley, 2004.
14. T. M. Connolly and C. E. Begg, **Database systems: a practical approach to design, implementation, and management**, 6. ed., global ed. Boston, Mass.: Pearson, 2015.
15. J. W. Satzinger, ***Systems analysis and design in a changing world***, 7th edition. Boston, MA: Cengage Learning, 2015.
16. C. Vercellis, ***Business intelligence: data mining and optimization for decision making***. Chichester, U.K: Wiley, 2009.
17. W. H. Inmon, ***Building the data warehouse***, 4th ed. Indianapolis, Ind: Wiley, 2005.
18. L. Yessad and A. Labiod, **Comparative study of data warehouses modeling approaches: Inmon, Kimball and Data Vault**, in 2016 International Conference on System Reliability and Science (ICSRS), 2016, doi: 10.1109/icsrs.2016.7815845.
19. S. Vyas and P. Vaishnav, **A comparative study of various ETL process and their testing techniques in data warehouse**, Journal of Statistics and Management Systems, vol. 20, no. 4, pp. 753–763, Jul. 2017, doi: 10.1080/09720510.2017.1395194.
20. R. Kimball and M. Ross, ***The data warehouse toolkit: the definitive guide to dimensional modeling***, Third edition. Indianapolis, IN: John Wiley & Sons, Inc, 2013.
21. S. Scheps, ***Business intelligence for dummies***. Hoboken, NJ: Wiley, 2008.
22. A. S. Girsang et al., **Decision support system using data warehouse for hotel reservation system**, in 2017 International Conference on Sustainable Information Engineering and Technology (SIET), 2017, doi: 10.1109/siet.2017.8304166.
23. M. Yulianto, A. S. Girsang, and R. Y. Rumagit, **Business intelligence for social media interaction in the travel industry in Indonesia**, Journal of Intelligence Studies in Business, vol. 8, no. 2, Sep. 2018, doi: 10.37380/jisib.v8i2.323.
24. K. Morton, R. Bunker, J. Mackinlay, R. Morton, and C. Stolte, **Dynamic workload driven data integration in tableau**, in Proceedings of the 2012 International conference on Management of Data - SIGMOD '12, 2012, doi: 10.1145/2213836.2213961.