Aishwarya Iyengar | Smita Kulkarni | Swathi Badicole

Analyze Movie Big Data Set for Crowdfunding

2021

# Introduction

As a part of this Project, we will be dealing with the business analytics of Movie data set to support crowdfunding decisions.

Problem Statement:

In the current world, crowdfunding for an independent film is one of the most popular form of alternative finance. Crowdfunding in simple words is asking a crowd for your movie funding. But, in order to fund any movie, a good data is critical. Many investors who fund for a movie need data on the directors or actors or genre past performance. They need data to understand what kind of movie is doing really well and what kind of movie is not. These kind of analysis on any movie dataset can help with crowdfunding. And there is no good project today supporting these kinds on analysis which will help both the movie maker to set up a campaign based on his past records and investors who want to decide with whether to proceed with funding a particular movie or not.

Goal:

1. Understand the movie review dataset (IMDB and Movie Industry)
2. Understand and perform the ETL process required on the movie review datasets using amazon glue
3. Understand data-warehouse and the process required to transform the data.
4. Understand how Amazon Redshift can help with data warehousing.
5. Understand how business analytics can be performed on Amazon redshift data-warehouse using Amazon QuickSight(Or other BI tools)
6. Discuss how analytics done from our transformed data can help with crowdfunding decisions

Our Audience:

1. Investors/ Funders to movie
2. Movie makers interested in raising a crowdfunding campaign
3. General Public interested in movie trends

Why this dataset?

* This real-time dataset from IMDB and Movie Industry have the data of movies starting from 1874-2021 across various countries
* Movie dataset from IMDB was supporting our analysis from crowdfunding perspective

How will this help?

This analysis can help businesses understand

* This analysis can help general investors/public/movie makers interested in crowdfunding to understand
* What genre movies are popular
* What movies have highest amount of ratings
* This analysis can be a pre-requisite for building a recommendation model to any investor/movie maker.

# Steps Planned to be followed:

1. To bring in Movie review data set into the Amazon’s S3 bucket. Our data source is from two different sources as below:
2. IMDB Movie Review Dataset from IMBD

<https://www.imdb.com/interfaces/>

1. Movie Industry Data set from Kaggle

<https://www.kaggle.com/danielgrijalvas/movies>

1. Use Amazon Glue to read data from Amazon S3
2. Use PySpark (Python Code) on Glue Job Run functionality to perform any transformation on our dataset if required.
3. Load the extracted and transformed data into Amazon Redshift using Amazon Glue endpoint connection.
4. Run Data Analytics on Amazon Redshift Dataset. Example: Which director movie has maximum number of good reviews, Which genre movie is performing the best, which are the top movies with maximum number of reviews etc.

# Project Timeline & Plan:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Task | Assigned to | Start Date | Due Date | Meeting | Status |
| 1. | Project Plan and Proposal | Smita | 09-04-2021 | 10-04-2021 | 10-04-2021 | Done |
| Identify Data Source | Smita | 07-04-2021 | 09-04-2021 | 09-04-2021 | Done |
| 2. | Project Environment Set Up in AWS(End-to-End)   * S3 Bucket Creation * Data Load into S3 * Amazon Glue initial Setups * Create endpoints from Glue as required * Crawler Setup * Amazon redshift Set up | Smita | 10-04-2021 | 17-04-2021 | 17-04-2021 | Done |
| 3. | Extract and Transform (Glue Job)  Data-ware house Design  Data Cleaning  Data Preprocessing  Data from S3 to redshift | Smita | 17-04-2021 | 24-04-2021 | 24-04-2021 | Pending for Movie Industry Dataset |
| Project Report Documentation Start   * Abstract * Introduction * Related Work * Dataset Description * Problem Statement | Aishwarya | 17-04-2021 | 24-04-2021 | 24-04-2021 |  |
|  | Project Report Documentation Start   * Proposed Solution * Project Architecture Diagrams, Description and Implementation | Swathi | 17-04-2021 | 24-04-2021 | 24-04-2021 |  |
| 4. | Run Data Analytics on Amazon Redshift using BI Tools   * Different Visualization Graphs/Charts | Aishwarya | 24-04-2021 | 01-05-2021 | 01-05-2021 |  |
| Project Report Documentation Contd.   * Project Methodology/Deep Dive Details   PowerPoint Skeleton for presentation | Smita | 24-04-2021 | 01-05-2021 | 01-05-2021 |  |
| Project Report Documentation Contd.   * Results * Future Work and Scope * References * Report Documentation Close | Swathi | 24-04-2021 | 01-05-2021 | 01-05-2021 |  |
| 5 | PowerPoint Close | Aishwarya | 01-05-2021 | 04-05-2021 | 04-05-2021 |  |
| 6 | Rehearsal |  | - | - | 08-05-2021 |  |
| 7 | Project Presentation |  |  |  | 11-05-2021 |  |

# Initial Data Set Analysis:

IMDb Dataset:

1. title.basics.tsv.gz

Table

Description automatically generated

2) title.ratings.tsv.gz

Table

Description automatically generated

3) title.principals.tsv.gz

Table

Description automatically generated

4) name.basics.tsv.gz

Table

Description automatically generated

Movie Industry Dataset:

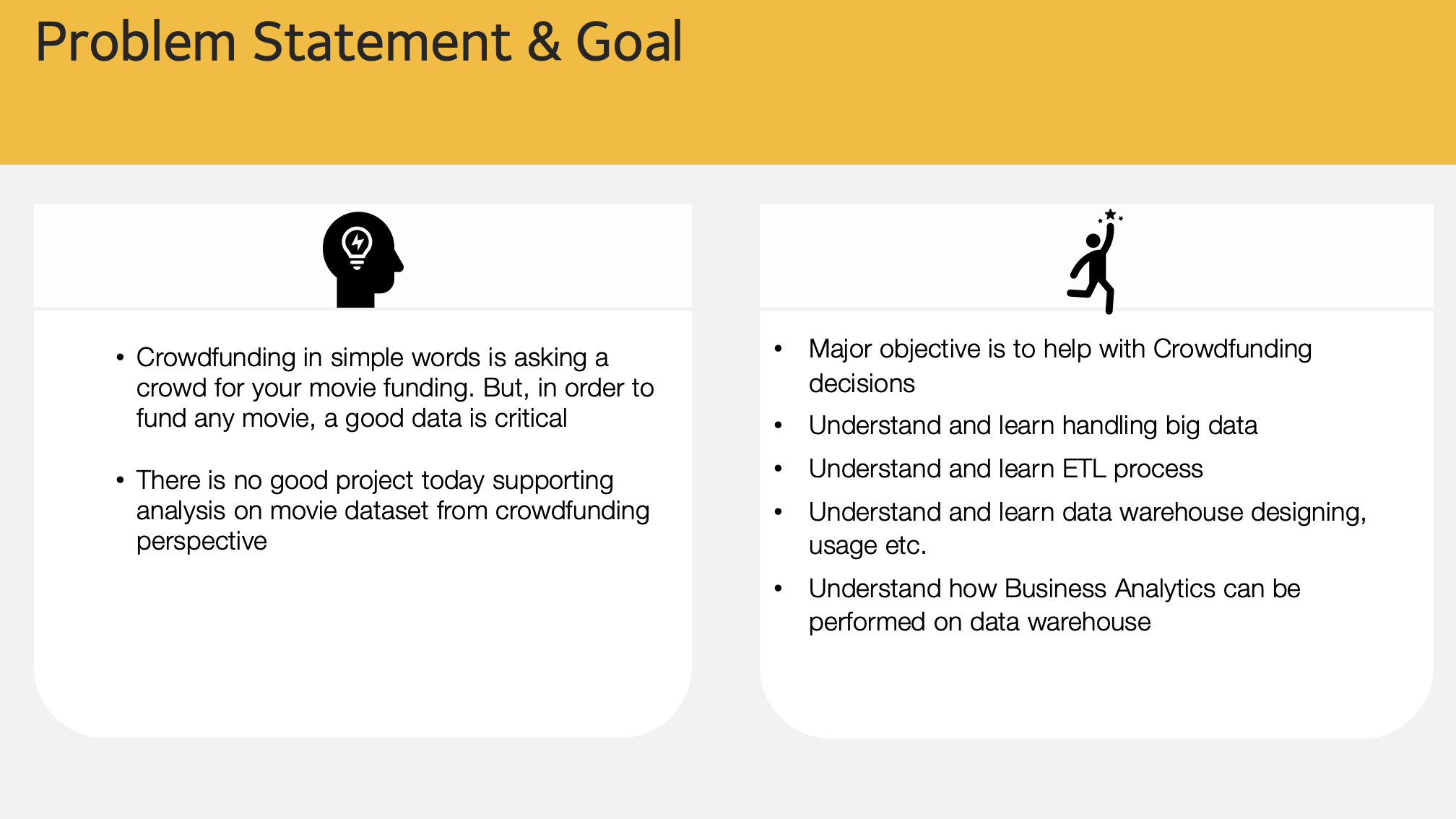
1)movies.csv

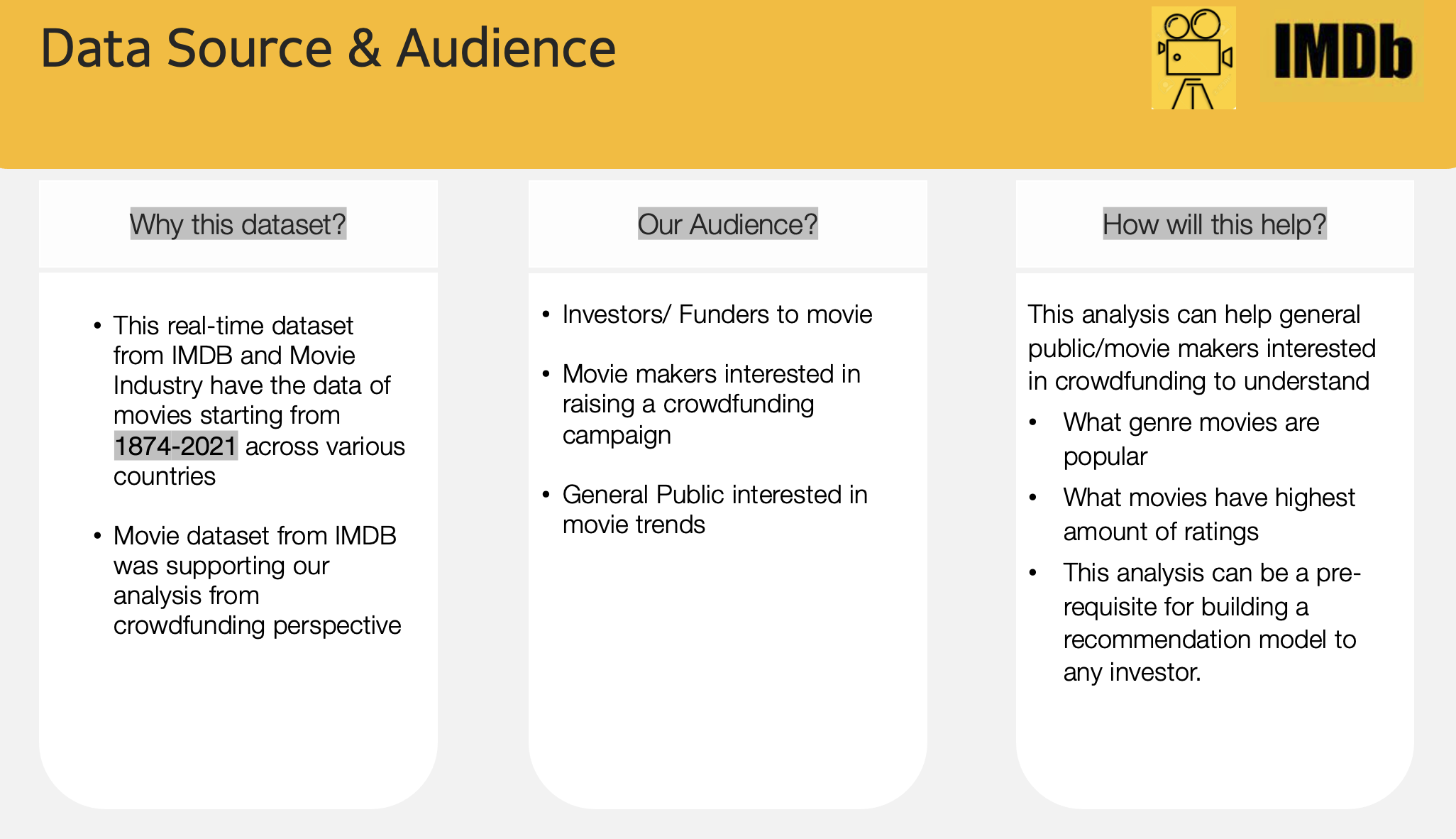
Table

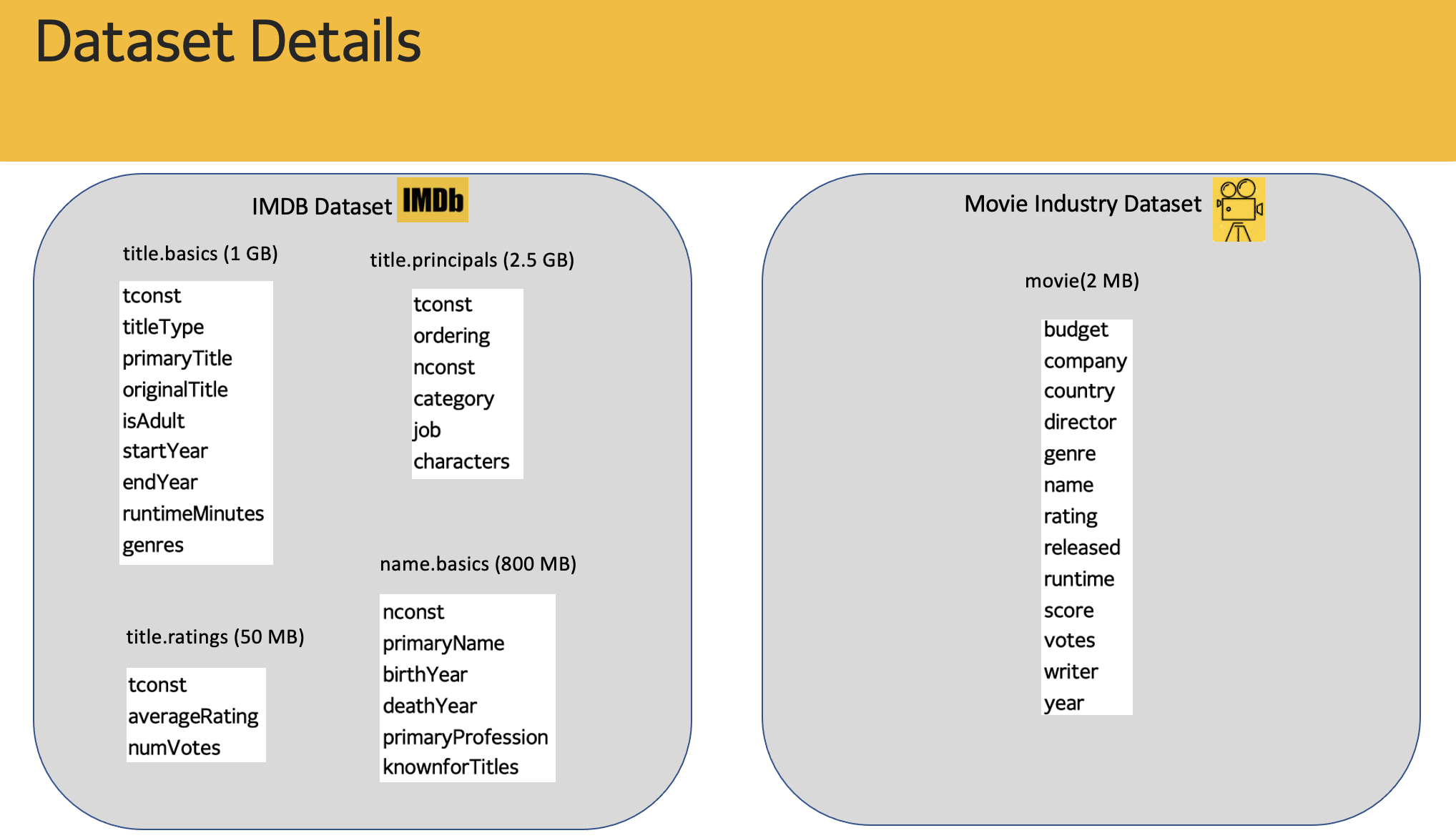
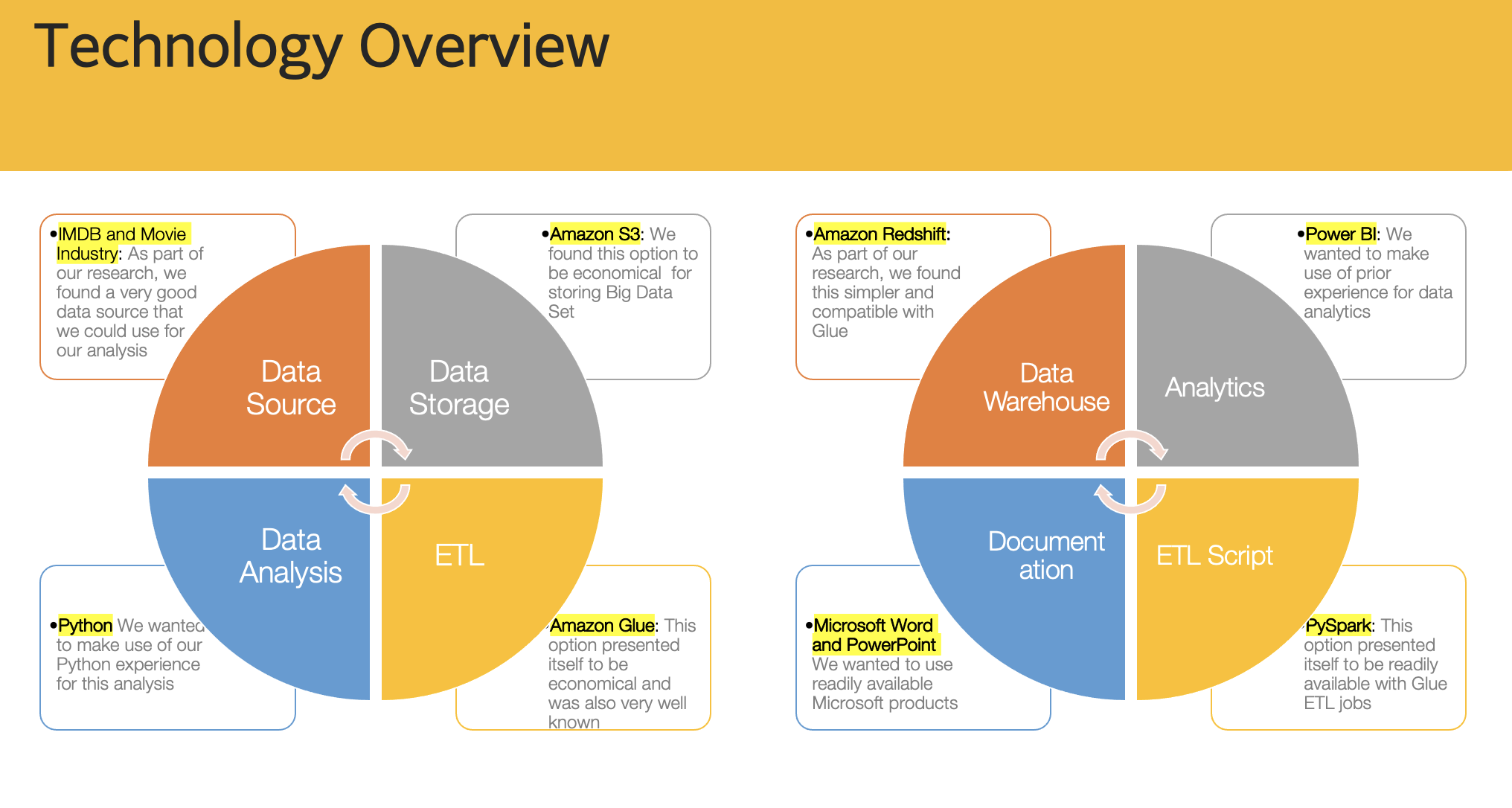
Description automatically generated

# Project Proposal PowerPoint



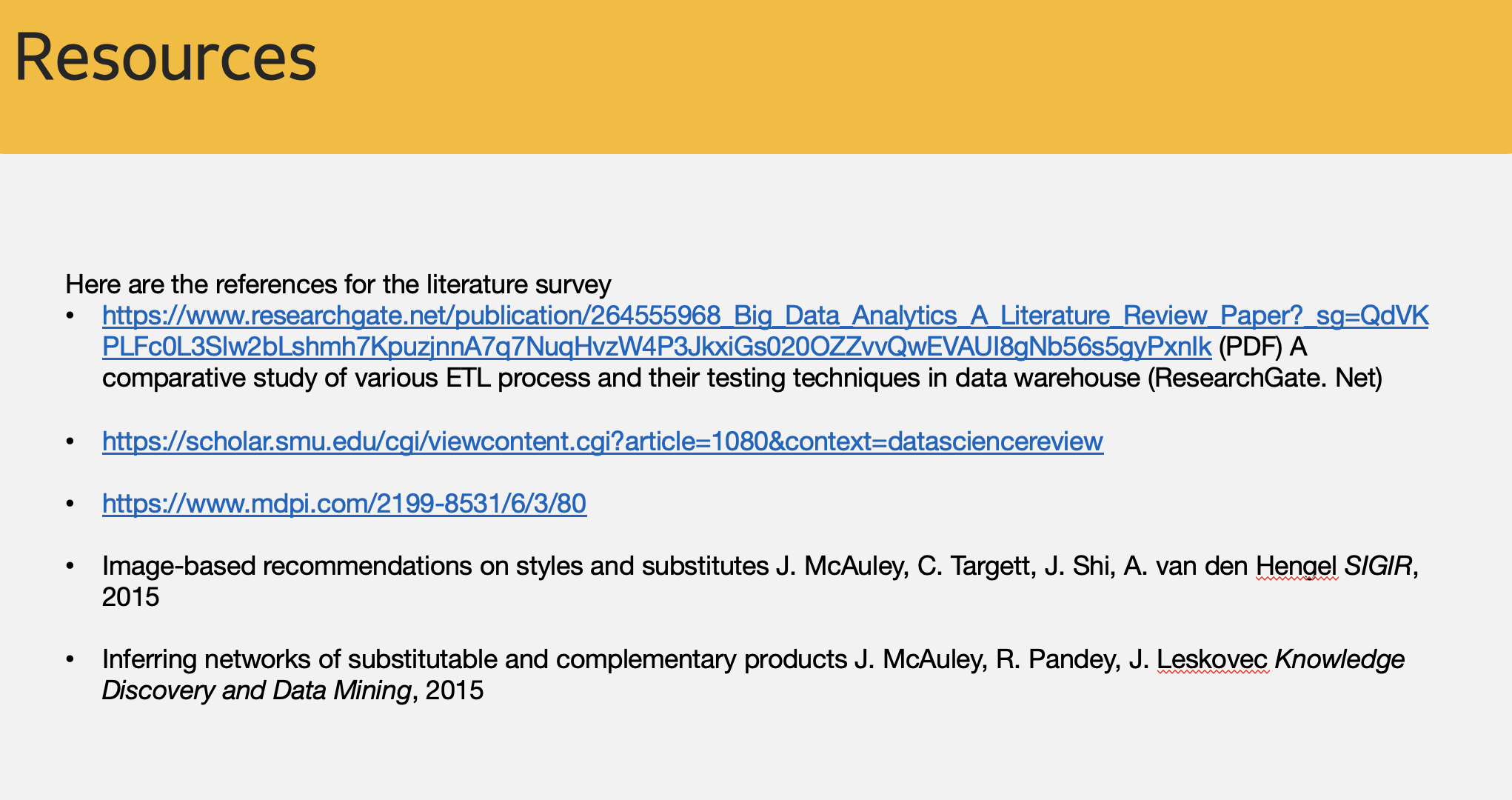




Graphical user interface, diagram

Description automatically generated



# Citation

DATASET

IMDB Movie Review Dataset from IMBD

<https://www.imdb.com/interfaces/>

Movie Industry Data set from Kaggle

<https://www.kaggle.com/danielgrijalvas/movies>