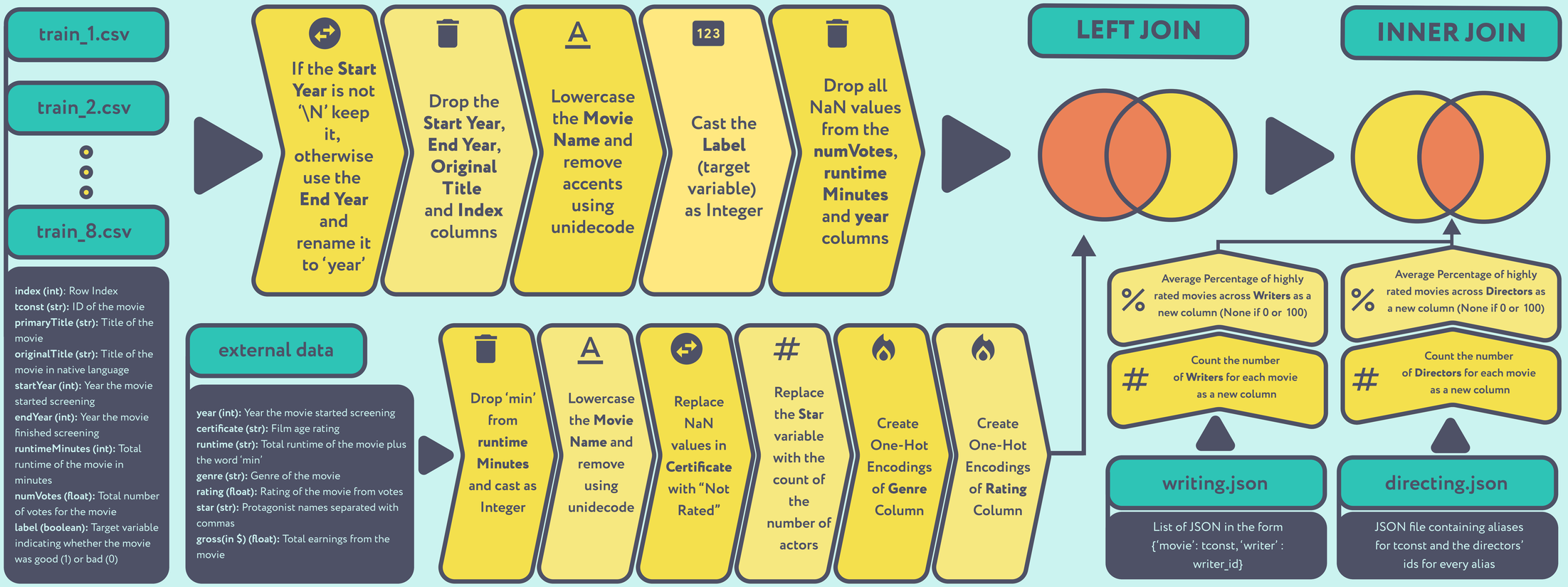
BIG DATA INSIGHTS: IDENTIFYING BLOCKBUSTERS TO FLOPS

Hendrik Jilderda, Alexander Koch, Ioannis Papageorgiou, Efstathios Trigkas, Dimitrios Tsiamouras



PROBLEM STATEMENT

- **IMDd Dataset:** prediction of highly rated movies using data such as titles, start and end dates, run time and votes, as well as data about the writers and directors.
- **PySpark** : chosen for its ability to **scale** efficiently without significant **execution speed** overhead.
- Incorporated additional **external data**, including movie age ratings, genres, ratings, actors and total earnings.
- Both datasets contained multiple **NaN** values, **redundant** columns, as well as data that needed to be **processed** to be utilized in an appropriate form
- **XGBoost**: chosen for the prediction task, as it is considered the state-of-the-art model for tabular machine learning tasks and can handle NaN values.

81.26%
ACCURACY

PARAMETERS:
Number of Estimators: 200
Max Depth: 3
Learning Rate: 0.15

dmlc XGBoost

CONCLUSIVE REMARKS

- **Two data pipelines** were developed for both the given data as well as the external dataset.
- The same pipelines were applied to the **validation** and **test** datasets without any complications.
- The **external data** increased the model's performance in conjunction with **fine-tuning** and resulted in an **accuracy of 81.26%** on the validation set
- **PySpark** showed **low execution times** and **efficiently processed** the combination of internal and external data..
- Additional external data along with feature engineering, such as **sentiment analysis** on **movie reviews** are expected to increase the model's performance further, while they could be efficiently handled using PySpark