From VFX Project Management to Predictive Forecasting

Hannes Ricklefs
MPC

Stefan Puschendorf
MPC

Sandilya Bhamidipati
Technicolor Research

Brian Eriksson
Technicolor Research

Akshay Pushparaja
Technicolor Research

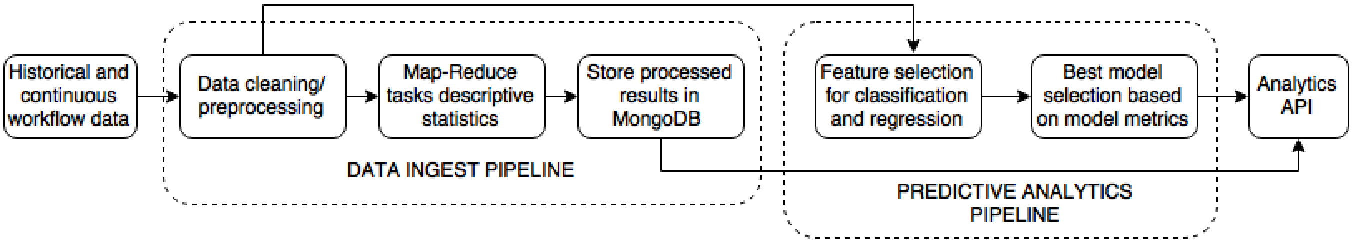
Historical and continuous workflow data
Data cleaning/ preprocessing
Map-Reduce tasks descriptive statistics
Store processed results in MongoDB
DATA INGEST PIPELINE
Feature selection for classification and regression
Best model selection based on model metrics
Analytics API
PREDICTIVE ANALYTICS PIPELINE

Figure 1: VFX Data Analysis and Predictive Analytics Workflow

ABSTRACT

VFX production companies are currently challenged by the increasing complexity of visual effects shots combined with constant schedule demands. The ability to execute in an efficient and cost-effective manner requires extensive coordination between different sites, different departments, and different artists. This coordination demands data-intensive analysis of VFX workflows beyond standard project management practices and existing tools. In this paper, we propose a novel solution centered around a general evaluation data model and APIs that convert production data (job/scene/shot/schedule/task) to business intelligence insights enabling performance analytics and generation of data summarization for process controlling. These analytics provide an impact measuring framework for analyzing performance over time, with the introduction of new production technologies, and across separate jobs. Finally, we show how the historical production data can be used to create predictive analytics for the accurate forecasting of future VFX production process performance.

CCS CONCEPTS

•Information systems →Enterprise resource planning;

KEYWORDS

Data Analytics, Machine Learning, Business Intelligence, VFX Workflows

ACM Reference format:
Hannes Ricklefs, Stefan Puschendorf, Sandilya Bhamidipati, Brian Eriksson, and Akshay Pushparaja. 2017. From VFX Project Management to Predictive Forecasting. In *Proceedings of SIGGRAPH '17 Talks, Los Angeles, CA, USA, July 30 - August 03, 2017,* 2 pages.

SIGGRAPH '17 Talks, Los Angeles, CA, USA

DOI: http://dx.doi.org/10.1145/3084363.3085036

DOI: http://dx.doi.org/10.1145/3084363.3085036

1 INTRODUCTION

Over the span of several years, MPC has built considerable internal software infrastructure around data reporting and project management [Ricklefs 2013]. These software systems track the progress and work of individual artists, account for working hours, and aggregate performance of work across multiple levels (*e.g.,* job, scene, shot). This data allows for department managers to better assess the progress of their teams. While this level of detail suffices for small projects, digging through all this data can become overwhelming when looking at larger projects or trying to aggregate across an entire organization. As MPC grows to multiple sites and thousands of employees, new tools become necessary to make sense and derive insights from this data.

The first effort was focused on business intelligence, specifically on implementing a real-time data pipeline to ingest production data and process relevant key performance indicator (KPI). Using a MapReduce framework, we constructed a cloud-based data ingest pipeline that takes project management data and outputs descriptive statistics. These statistics are structured into an *Evaluation Matrix* format that aggregates these statistics into a desired organizational level (*e.g.,* job, shot, scene). The goal is to use these statistics to create holistic insights into segments of the MPC business.

The second effort was to exploit this large corpus of historical project data to generate predictive models to forecast future project performance. This was implemented as a real-time predictive analytics pipeline that ingests current project data, trains machine learning models, and output predictive forecasts for each project group. With our initial models, we currently have a 73% accuracy rate when predicting if a specific task will overspend. These initial results reveal the promise of using predictive analytics for large-scale VFX production.

2 DATASET

The MPC Project Management dataset consists *of both continuous integration* of new production data and a large corpus of historical production data from January 2012 to October 2016. This historical data has over 1.3 million entries containing VFX disciplines, artists, sites, bids and actual times associated with the specific tasks.

3 DATA INGEST PIPELINE

In order to understand current production process performance and identify areas which can be improved for efficiency, we built the data processing pipeline shown in Figure 1 to measure overspend, overtime, performance and job duration. In the initial stage, data cleaning identifies valid shots and builds for further processing. The next stage creates summary and breakdown statistics for site-department-jobs-scene-shots over multiple time durations ranging from years to weeks. This is efficiently performed by passing the data to two separate multi-level MapReduce [Dean and Ghemawat 2008] tasks corresponding to the year time-scale and week time-scale, respectively. The final stage stores the statistics results in a MongoDB database [MongoDB 2017]. All three stages are repeated regularly as new data arrives, providing updated production statistics in real-time.

4 PREDICTIVE ANALYTICS PIPELINE

Using the predictive analytics pipeline, we estimate production task actuals using machine learning models trained from our historical data corpus. The overall predictive analytics pipeline can be seen in Figure 1.

To rigorously train a machine learning model using "cross-validation" [Friedman et al. 2001], we first split the cleaned data from the ingest pipeline into separate train and test sets. These train and test datasets are build based on the job completion status and the data timeline. Currently, all completed jobs from 2014 to 2015 are used for training and jobs from 2016 for testing. This component was specifically designed to be adaptable as new job data is added.

Next we translate the unstructured production data into a format that can be used in machine learning algorithms. Specifically, this requires each tasks to be reduced to a series of observed, relevant features in a vector format. We convert the id numbers for each artist, department, site, and job into a categorical binary variable. The end result is each task being represented by a feature vector of length X.

Using the structured train and test sets, we next identify a set of candidate classification or regression machine learning models for a given objective. Specific machine learning models used are off-the-shelf using Python Sklearn [Pedregosa et al. 2011] (*e.g.,* Ridge classifier, Nearest Neighbor Classifier, Gradient boosted classifier). All machine learning models have a set of parameters that must be learned via optimization on the training set, with subsequent accuracy measurement of the model on the test set. The best model corresponds to the model that provides the best accuracy metrics on the test set (*e.g.,* misclassification rate, F1 score, precision score).

Finally, we save the trained model parameters to be packaged in an API. As new data is made available, the Machine Learning (ML) pipeline is re-run with the updated test and train sets to retrain the machine learning models. The modular design of the ML pipeline gives numerous benefits,

- *(1)* - The ability to create new test and train models on the fly,
- *(2)* - the ability to add new custom features with minimum changes to existing code base, and
- *(3)* - and the ability to easily include new machine learning (or deep learning) models.

Initial results indicate the ability to accurately forecast future production performance. Our results show that gradient boosted classifiers have an accuracy score of over 73.25% on our test data for classifying whether a task will overspend on initial estimates. These initial results show the promise of using machine learning to predictive future VFX production process performance.

5 CONCLUSION AND FUTURE WORK

In this paper we presented a data analytics pipeline for analyzing VFX production data and providing predictive forecasts using machine learning. This work is a first attempt at creating intelligent tools to cope with the massive amounts of data generated by modern VFX organizations. Our initial results show that this framework provides key insights into production and obtains considerable accuracy with respect to common production problems (*e.g.,* Bid vs Actuals overspend). Next steps include providing more comprehensive business insights by incorporating additional datasets (*e.g.,* asset management, resource planning, render and data consumption) to fully understand and evaluate the complete production process. Additionally, we look to further refine and add new machine learning algorithms to our predictive analytics pipeline.

REFERENCES

Jeffrey Dean and Sanjay Ghemawat. 2008. MapReduce: Simplified Data Processing on Large Clusters. *Commun. ACM* 51, 1 (jan 2008), 107–113.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. 2001. *The Elements of Statistical Learning.* Vol. 1. Springer Series in Statistics.

Inc. MongoDB. 2017. *MongoDB.* http://www.mongodb.com/.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

Hannes Ricklefs. 2013. Pronto: Scheduling the Un-schedulable. In *ACM SIGGRAPH 2013 Talks (SIGGRAPH '13).* ACM, New York, NY, USA, Article 29, 1 pages. DOI: http://dx.doi.org/10.1145/2504459.2504495