Evaluating Dependability and Performance for Machine Learning at Scale

Donna Xu (Dr. Liming Zhu (External), Dr. Sanjay Chawla (Internal))
School of Information Technology, National ICT Australia
FACULTY OF ENGINEERING & INFORMATION TECHNOLOGIES

BACKGROUND AND MOTIVATION

Big Data

- High volume and high velocity
- There are many research works show that their machine learning algorithms are developed for running on a single machine [1, 2], which have limitation on handling large amount of data
- Machine learning developers need to migrate and deploy Machine Learning (ML) programs from single machine environment to distributed environment
- High velocity of big data needs to be handled for real time predictive analytics

Online Machine Learning Service

- Wrap machine learning as a service which makes machine learning model accessible to people who are not familiar with it
- Integrate learned models with real time analysis and prediction services
- Update learned models with streams of new data

Dependability

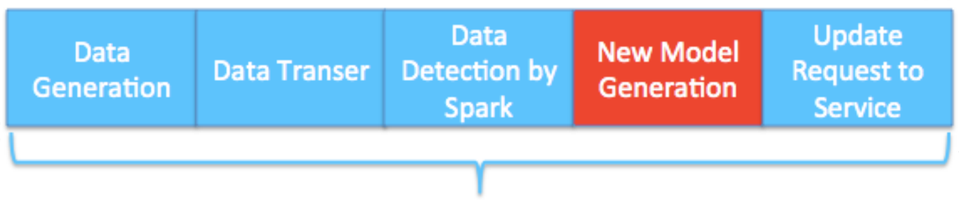
- Frameworks such as MLBase/Spark [3, 4] and Mahout/Hadoop [5, 6] help machine learning developers and end users, there are still dependability and performance issues at both infrastructure level and application level

OBJECTIVES

- Design and implement RESTful training and prediction ML services for real time predictive analytics
- Evaluate its performance and dependability issues

SERVICE DESIGN

REAL-TIME ANALYTICS LATENCY



LATENCIES

- Data generation – can not be controlled by our system
- Data transfer – latency depends on internal network delay
- Data detection by spark – latency depends on internal implementation of Spark and network delay
- New model generation – we are focusing on investigate the factors that affect the latency of model generation
- Update request to service – normally this is very small (< 0.08s) which can be neglected

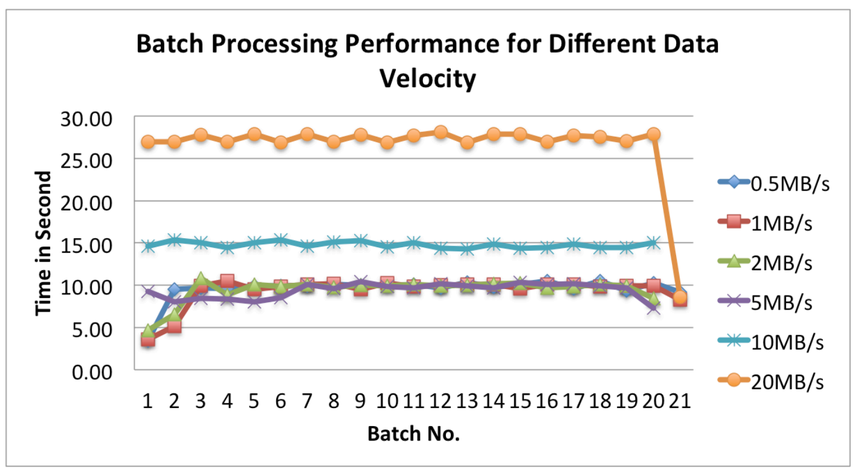
EXPERIMENT DESIGN & RESULTS

Data Volume

- Training historical data on different sizes

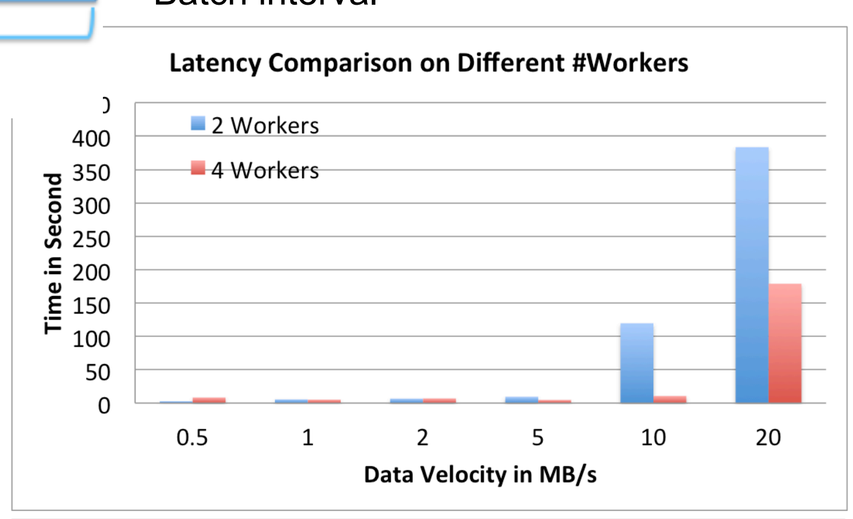
Data Velocity

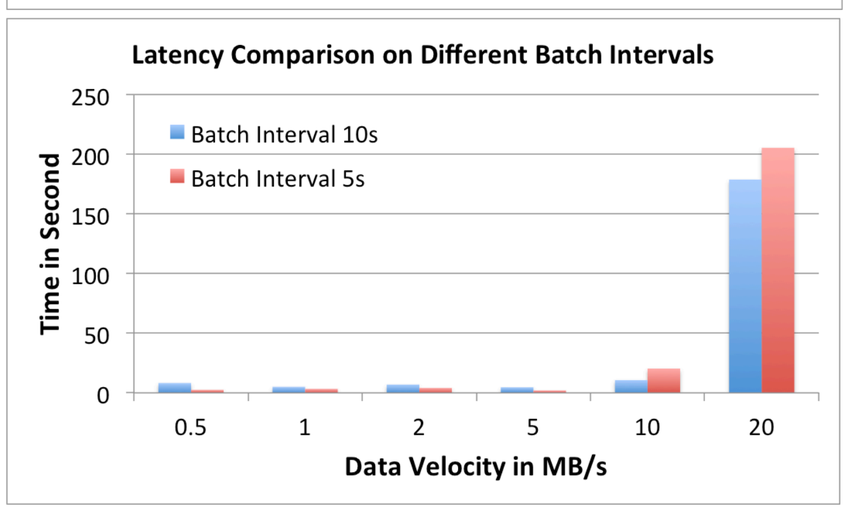
- Simulate data streaming to S3 bucket against different velocities
- Streaming input data for a continuous 200 seconds for each data velocity



Real-time analytics latency is affected by three factors

- Number of workers
- Data velocity
- Batch interval

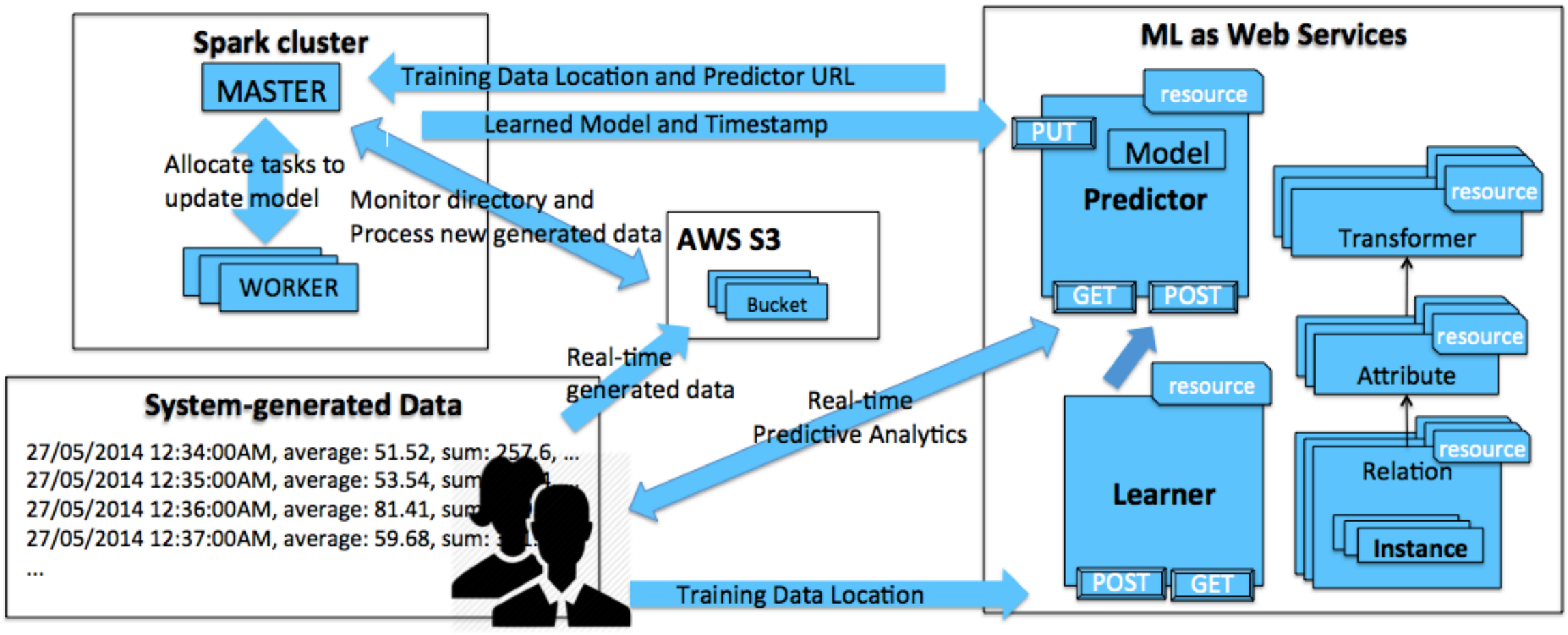




EVALUATION

Performance and Dependability

- The system is able to handle up to 5MB/s live streaming data using two workers and 10MB/s data using four workers with batch interval 10 seconds
- Highly unreliable to train 32GB historical data using four workers
- Latency can be reduced by increasing the number of workers
- Setting batch interval to 5 seconds performs better in real-time analysiss with velocities < 5MB/s than setting to 10 seconds



REFERENCES

[1] Xu, Wei, et al. "Detecting large-scale system problems by mining console logs." Proceedings of the ACM SIGOPS 22nd symposium on Operating systems principles. ACM, 2009.
[2] Fu, Qiang, et al. "Execution anomaly detection in distributed systems through unstructured log analysis." Data Mining, 2009. ICDM'09. Ninth IEEE International Conference on. IEEE, 2009.
[3] MLBase. http://www.mlbase.org.
[4] Spark. http://spark.apache.org.
[5] Apache Mahout. http://mahout.apache.org.
[6] Apache Hadoop. http://hadoop.apache.org.