A Generalized Scalable Software Architecture for Analyzing Temporally Structured Big Data in the Cloud

Magnus Westerlund, Ulf Hedlund, Göran Pulkkis, and Kaj-Mikael Björk

Arcada University of Applied Sciences, Jan-Magnus Janssons plats 1,
00550 Helsinki, Finland
{magnus.westerlund,ulf.hedlund,goran.pulkkis,
kaj-mikael.bjork}@arcada.fi

Abstract. Software architectures that allow researchers to explore advanced modeling by scaling horizontally in the cloud can lead to new insights and improved accuracy of modeling results. We propose a generalized highly scalable information system architecture that researchers can employ in predictive analytics research for working with both historical data and real-time temporally structured big data. The proposed architecture is fully automated and uses the same analytical software for both training and live predictions.

Keywords: predictive analytics, temporal data, system-level design, self-adaptive systems, runtime models.

1 Introduction

Development of big data analytics enabled solutions are being driven by unstructured data, surprisingly little is done in open sourced development of processing structured big data. Structured data is often considered, mischievously so, as data at rest perhaps due to its usual implementation as a state storage. In the world of big data analytics, it is not enough to simply analyze historical or at rest data, instead this must be done in a (near) real-time environment. The term real-time environment means something slightly different depending on the context, here we refer to an event being triggered when new data exists and how the system consequently handles that data automatically. We concur with the argument that a big data model's excellence can only be shown after it has been employed in an online environment [1][2]. In [3] it is claimed that "Advancing the cloud from a computation and data management infrastructure to a pervasive and scalable data analytics platform requires new models, tools, and technologies that support the implementation of dynamic data analysis algorithms." The process of integrating disparate data sources (external and/or internal origin), continuously preprocessing data for validity or other purposes, and transporting data to processing nodes is such a delicate and often limiting step in the workflow that a model's performance can only be known after this step has been completed. Quite few predictive models exist that have an ability for online learning; this adds a

Á. Rocha et al. (eds.), *New Perspectives in Information Systems and Technologies, Volume 1*,
Advances in Intelligent Systems and Computing 275,
DOI: 10.1007/978-3-319-05951-8_53,

requirement for models employed in a live environment to be re-trained if/once their effectiveness diminishes. This self or auto tuning is an integral part of any real-time automated analytical system.

We propose a generalized scalable software architecture based on predictive analytics for working with structured (near) real-time and historical data. The structured data we refer to can originate from a multitude of sources like sensor data, machine or user interface, financial data or any other source of online data with a temporal structure, commonly referred to in programming terms as data tuples. The proposed architecture will be able to take advantage of resources from both a public cloud infrastructure and private cloud providers [4]. We set a requirement that data processing activities should be fully automated once the system user has initiated model training. The architecture should automatically handle scaling on available resources, perform feature extraction, carry out model training, initiate auto tuning, and do live prediction.

We focus on the information system architecture and make the assumption that global memory is not usually required for analytical models that consume temporally structured big data. The contribution we bring is an understanding of a generalized light-weight scalable architecture for designing actual analytical information systems that can be used in collaboration with modern decision support systems. Our architecture allows the researcher to take almost any off-the-self model that can be encapsulated as an executable binary and deploy it as a scalable predictive analytics information system. By making extensive use of cloud computing in order to be able to scale sufficiently, we can employ an ensemble of ensembles of different machine learning models to achieve a good prediction rate [5]. We primarily narrow the type of forecasting models we use to those using supervised learning, e.g. Support Vector Machines [6] or Neural Networks [7]. This does not however exclude the use of unsupervised or reinforcement based learning methods provided a model and training error calculation can be automatized.

2 Related Research

Analytical Information Systems have recently received a great deal of attention from academics, open source community, as well as industry [8][9][10][11][12]. Also standardization steps in big data analytics have been taken [13]. The driving force behind related software design has been the Service Oriented Architecture (SOA) [14] approach, which has become an industry de-facto standard for building cloud based, loosely-coupled “X as a Service” enabled data sharing software modules.

The open source community tools for processing and storing big data [15], e.g. technologies such as Hadoop [16] and PIG [17], are mainly focusing on unstructured data and on extending the MapReduce programming model [18]. The MapReduce programming model assists the developer in segmenting data into smaller pieces and process them on the node were data resides, instead of moving data to a second node for processing. This consequently reduces processing time for large data sets.

In the case of handling structured big data, as temporal data often is, more traditional methods are still often applied, such as queuing methods and clustered