A General Scalable Framework for Integrating Big Data Programming Models with Unified Abstraction (PI: Prof. Cao Jiannong; 2013/14)

Big data is characterized by its volume, velocity and variety, and applies to information that exceeds the processing capacity of conventional database systems [56]. Cloud computing provides abundant computing resources and massively parallel processing capabilities that can support the management and processing of big data. In the recent years, many projects have been carried out on coordinating the cloud computing resources to tackle the big data problem. In particular, programming models have been proposed to support the programming of different types of big data applications.

In this project, we study the problem of providing high-level programming support of big data applications that can accommodate programming models for different application requirements. Although there exist various programming models and tools for processing different types of data, for example, MapReduce model for batch data processing [4], Pregel for graph processing [5], and S4 for stream data processing [6], no single model can support the diverse requirements and types of big data applications. There are some platforms that provide integration of the models, but they require the big effort of programmers to specify the applications mapping into separate underling models. The existing models deployed on the cloud require the administrators to manually configure the resources in advance, but the required amount of resources dynamically changes depending on the stages of the data processing.

This project aims to develop a general scalable framework for integrating big-data programming models with a novel approach to providing unified abstraction for programming various big data applications. The meaning of ‘scalable’ is twofold: allowing the addition of new programming models, and elastic resource management

for flexible data valume. We have investigated that none of the existing works has achieved the objectives.

Many challenging issues need to be addressed. First, since existing big data programming models have different design, it is not easy to integrate them and provide a unified programming interface. Second, it is challenging to design the mechanisms to abstract the underlying execution heterogeneity of various programming models. Third, in order to support dynamic utilization of the cloud resources, the framework needs to accurately analyze the types of computations in the application. We will first investigate the requirements of big data processing, and categorize the existing big data programming models based on the data types and operations they support. Second, we will develop a general approach to expressing different operations on various types of data, and implement a general interface for programming big data applications. To support the general interface, we will design middleware which provides functions to bridge the execution heterogeneities of various models. The functions include mapping various stages of applications onto the invocations of operations supported by different models, scheduling various jobs/tasks from the model invocations onto the resources, analyzing the required computations of the application and automatically configure the resources from clouds.

This project will make significant contribution for programming big data applications, which benefit business, industry, and the society in discovering knowledge and making decisions. The proposed research will help ease and shorten the development of big data applications by providing a unified programming interface. Also, our framework takes advantages of the 'pay-as-you-go' flavor cloud resources, so it can also help the developers save cost.