Bring Big Data To Everyone



1. Description of research project

1.1 Introduction

In recent years, researchers invented many distributed programming frameworks to handle large scale machine learning tasks. In the meantime, we are entering a big data era. Users from different areas are all demanding to understand the large volume of data. However, most of the users cannot handle the complexity brought by the large volume of data, and also the machine learning (ML) algorithms. This problem exacerbates when it is handled in the distributed system environment. Even for many of machine learning experts without a strong background in distributed systems, they cannot use the existing frameworks to implement algorithms efficiently.

In this project, we aim to develop a powerful and user-friendly system to lower the barrier to conduct large-scale machine learning tasks. Our main goal contains two main parts:

1. Implement a scalable system to handle large input set;
2. Provide user-friendly environment to mask the complexities of algorithms and distributed systems

1.2 Research Goals

1.2.1 Implement a scalable system to handle large machine learning task

(i) Machine learning computing model.

There are common computations for many of the typical machine learning algorithms, such as gradient descending, matrix calculation, etc.

In this subproject, we will develop a framework to handle these common tasks in ML with a fast, scalable and fault-tolerant manner. We provide these services as the basic primitives for users to implement the computational tasks efficiently.

(ii) Unified runtime environment for different data processing tasks.

Batch and stream processing handle historical and real-time data respectively. In today’s solutions, batch and stream processing are independent systems.

In subproject, we will develop a unified runtime environment to:

1. Improve the consistency and fault-tolerance of stream processing;
2. Design new batch processing engine supporting (approximate) real-time queries;
3. Design fast and accurate mechanism to sync real-time and historical data.

1.2.2 User Friendly Environment

(i) Unified Data Accessing Interface

Data can exist in diverse data management systems within a single organization, introducing multiple sets of APIs. It's not practical to expect a common typical user to master all APIs.

We aim to develop a unified interface which supports accessing data from different data sources, so that users do not need to bind themselves to a certain data management solution and struggle in learning different APIs.

(ii) Multiple-level abstraction to describe machine learning task workflow

Existing large-scale computing frameworks only provide unique abstraction to describe computation tasks.

Our goal is to provide multiple level abstractions to describe the workflow of machine learning algorithms. For users with limited technical background, we can provide them with high-level description language to run mature learning models on the dataset. For machine learning researchers, we provide operator-level abstraction for them to build ad-hoc algorithms.

(iii) Automatic Optimization Engine

The user-friendly optimization engine should make the following optimization automatically:

1. Data layers for intermediate results: The intermediate results which may be involved in multiple iterations should be stored in memory to avoid unnecessary IOs.
2. Search space tailoring: The optimizer should be able to estimate the model learning time and the expected quality for every candidate configuration to avoid exploration in a large parameter space.
3. Amazon EC2 Usage Plan

To validate our proposed system, we plan to implement it on Amazon Elastic Cloud Compute. Our prototype will target on implementing the most popular machine learning algorithms with our system. The requested grant for two-year complimentary access of Amazon EC2 is expected to cover the expenses for our system implementation and technique validation. Specifically, we plan to

1) Utilize the Amazon EC2 as the testbed for development and debugging;

2) Conduct performance analysis and evaluation to validate the efficiency and effectiveness of our design.

We will publish our results on top conferences and journals and further apply external research fundings. Amazon’s generous funding support is greatly appreciated and will be acknowledged on our project website and also in our publications.