PROJECT PROPOSAL

[Project Name]

# Summary

This project is about implementing a text processing system, which is based on Hadoop framework and going to run on the AWS cloud platform, and evaluating this system, with running it on different AWS mode.

# Problems

Data sets are growing rapidly in part because they are increasingly gathered by cheap and numerous information-sensing mobile devices, aerial (remote sensing), software logs, cameras, microphones, radio-frequency identification (RFID) readers and wireless sensor networks.

Relational database management systems and desktop statistics and visualization packages often have difficulty handling big data. The work instead requires "massively parallel software running on tens, hundreds, or even thousands of servers".

With its distributed processing, Hadoop can handle large volumes of structured and unstructured data more efficiently than the traditional enterprise data warehouse.

## Previous Research

In October 1997, Michael Cox and David Ellsworth publish “Application-controlled demand paging for out-of-core visualization” [1]. They start the article with “Visualization provides an interesting challenge for computer systems: data sets are generally quite large, taxing the capacities of main memory, local disk, and even remote disk. We call this the problem of big data. When data sets do not fit in main memory (in core), or when they do not fit even on local disk, the most common solution is to acquire more resources.” It is the first article in the ACM digital library to use the term “big data.”

* + 1. Relational Database System

/TODO…

# Objectives

Our goal of this project is to show the impact and improvement of performance with a new design of computer architecture compared with a traditional legacy architecture.

Specifically, we will run our text processing program on three different platforms, which are AWS cloud platform with two instances, Hadoop pseudo-distributed mode with single machine, and raw Microsoft Windows with single machine. We will monitor and record the performance data while running. Then, we will analysis this data and output the final result.

# Procedures

This project can be divided into XXXX main steps, which are:

1. Build Hadoop development environment on different platforms mentioned above.
2. Grab some data from the Internet for processing.
3. Store the data on Hadoop Distributed File System (HDFS).
4. Design and implement the text processing program using the map-reduce technique.
5. Deploy the program on different platforms.
6. Run program and monitor the performance for several times.
7. Analysis the performance data.

# Research Methodology

Here are some techniques used in the project:

## 5.1 Apache Hadoop

Apache Hadoop is an open-source software framework written in Java for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware. All the modules in Hadoop are designed with a fundamental assumption that hardware failures are common and should be automatically handled by the framework.

Hadoop Distributed File System (HDFS) – a distributed file-system that stores data on commodity machines, providing very high aggregate bandwidth across the cluster;

Hadoop MapReduce – an implementation of the MapReduce programming model for large scale data processing.

## 5.2 MapReduce

MapReduce is a programming model and an associated implementation for processing and generating large data sets with a parallel, distributed algorithm on a cluster. Conceptually similar approaches have been very well known since 1995 with the Message Passing Interface standard having reduce and scatter operations.

A MapReduce program is composed of a Map() procedure (method) that performs filtering and sorting (such as sorting students by first name into queues, one queue for each name) and a Reduce() method that performs a summary operation (such as counting the number of students in each queue, yielding name frequencies). The "MapReduce System" (also called "infrastructure" or "framework") orchestrates the processing by marshalling the distributed servers, running the various tasks in parallel, managing all communications and data transfers between the various parts of the system, and providing for redundancy and fault tolerance.

## 5.3 Amazon Web Services

Amazon Web Services (AWS), is a collection of cloud computing services, also called web services, that make up a cloud-computing platform offered by Amazon.com. These services operate from 12 geographical regions[2] across the world. The most central and well-known of these services arguably include Amazon Elastic Compute Cloud, also known as "EC2", and Amazon Simple Storage Service, also known as "S3". Amazon markets AWS as a service to provide large computing capacity more quickly and more cheaply than a client company building an actual physical server farm.

## 5.4 Linux System

A Linux-based system is a modular Unix-like operating system, deriving much of its basic design from principles established in Unix during the 1970s and 1980s. Such a system uses a monolithic kernel, the Linux kernel, which handles process control, networking, access to the peripherals, and file systems. Device drivers are either integrated directly with the kernel, or added as modules that are loaded while the system is running.

# Key Personnel

Yifan Xu yxx152530

Reference:

https://www.evl.uic.edu/cavern/rg/20040525\_renambot/Viz/parallel\_volviz/paging\_outofcore\_viz97.pdf