# Introduction

Data management has a great importance for most of the companies to maintain their business. Since technology is developing and companies are growing every day, it creates a demand to manage a huge amount of data. Managing huge amount of data is still an on-going research field and engineers are working on this field to provide secure and reliable systems with faster response time to people. One of the methods to store the data is data distribution which distributes the data on different servers to manage the data. Data encryption and data replication can be used together with data distribution to ensure the secure management of distributed data and the reliability of the system.

Our project, CrypDist, aims for maximizing the performance of the delivery of data and providing fast access for the users by using a decentralized distribution system.

## Purpose of the System

The system is designed to implement a solution for providing ease of access to the genotype and phenotype data for researchers to enable them to conduct advanced research by analyzing and updating the data. This will be achieved by delivering a software system which will interact with the users to perform the operations on data. The system has a distributed database called “blockchain” to ensure the safety and reliability of continuously growing data.

## Definitions, Acronyms, Abbreviations

|  |  |
| --- | --- |
| HTTP | Hypertext Transfer Protocol |
| HTTPS | Secure Hypertext Transfer Protocol |

## Overview

CrypDist is a blockchain based distribution application that will be created with Java on the both client and server side. It is also going to use Akamai to handle the distribution part as an off the shelf component. CrypDist is going to use NetStorage facility of Akamai to store the raw data and it is going to distribute it with the help of Akamai Edge servers.

On the client side, CrypDist will be able to provide a summary data in form of a blockchain so that the user will be able to understand the content of the raw data itself without looking at the exact data. Blockchain structure is a data structure used for storing continuously growing list of records without tampering or revision. Hence, by the help of blockchain structure, CrypDist will be able to maintain the data dynamically and the synchronization will be handled automatically. Clients will be able download whatever part of the data they need, manipulate it and upload it back to servers to be reached in future as a new version. However, the data in the blockchain cannot be manipulated by anyone. In case of data addition, a block containing the abstract information will be added synchronously to blockchain on clients and the actual data will be distributed among servers. Thus, the encryption of the data will be already done by blockchain meaning that the user will never know where the data sent but he is always going to have the information in his own summary, blockchain. However, the raw data will already be open to anyone who has the blockchain. Here, CrypDist classifies the raw data to two different classes; one will include the genomics data which will be open and the second will include phenotype of the owner of the data which will be kept encrypted because it contains sensitive personal data. Thus, when a client wants to reach phenotype data, he will need be authenticated by our services.

The main functionality of CrypDist is that it avoids the centralized database architecture so that we can have a democratized data distribution system which avoids any block and always up to deliver the data as possible as quickly thanks to Akamai services. Our first and main focus is to manage genomics data since demand to work on that specific data is huge such that many researchers are working to avoid future possible mutations by examining the outputs of the past research.

The problem which CrypDist is planning to solve is currently tried to be handled with dbGaP system, which provides centralized database of Genotypes and Phenotypes. Using dbGaP has several disadvantages: Firstly, using one centralized server for taking data will be slow, when number of clients and size of the data is considered. Secondly, dbGaP stores all the data secure and getting access to this data requires several diplomatic steps, which takes very long time. However, the accessibility of Genotypes data does not need to be controlled because it does not contain any personal information and it is needed for statistical data. As a result, currently researchers are having problem for accessing the secure Genotypes data, which should have been public. Lastly, since the data is stored in centralized database, crashes, in-tolerated faults or political issues (e.g. shut down applications by governments) on the servers may cause loss of data, which the Wellcome Trust users had encountered in the U.K.

Previously, blockchain structure was used in genomics for distributing CPU load between machines by exchanging bitcoins with process power (<http://dnadigest.org/a-new-multi-centralized-cryptocurrency-coinami/>). However, CrypDist is the first application using blockchain for data distribution.

# Current Software Architecture

The blockchain is the key element of the digital currency system bitcoin. The blockchain is a distributed database. Bitcoin has a decentralized system because blockchain stores the data across its nodes of network. Bitcoin uses blockchain to record the bitcoin transactions between users. Network of communicating nodes provide maintenance (replication of the transactions and sending the transaction information to other nodes) for the blockchain. When a new transaction is added to a node, this transaction is broadcasted to other nodes. Each node in the network stores its own copy of blockchain. Although bitcoin uses blockchain to store the information about transactions, there is no current system keeping abstract data of human genome in its blockchain.