**ABSTRACT**

In order to effectively extract knowledge and analyze data, a data scientist/engineer needs clean and consistent data. Manipulating data into a clean, consistent state is commonly painstaking and arduous. This process typically consumes 80% of a data scientist’s time. While some solutions, such as Google Refine, aid in this objective, they do not handle large (>1TB) data sets well or at all.

Colossus is a tool that profiles, visualizes, and cleans large dirty datasets of different formats. In addition, it is able to merge and normalize data sets into one consistent data set.

**ARCHITECTURE**

1. Remote Shell (SSH)
2. Submit Colossus Job (Spark Job Submission)
3. Retrieve Results (SFTP File Transfer)
4. View Visualizations (HTTP Request)

How It Works:

Built on Apache Spark and Hadoop Distributed File System (HDFS), Colossus requires an active Spark cluster and HDFS cluster to run. Colossus consists of a set of Python command line scripts on an edge node, a computer on the same network as the HDFS and Spark clusters, and a set of Spark applications that run on the cluster.

The Python script submits the Spark applications to the Spark cluster which then executes those applications on datasets in the HDFS cluster. Upon completion, the script retrieves and renders the application results on a web server.

**SUMMARY**

Colossus builds upon mainstream big data tools Spark and the Hadoop Distributed File System to help data scientists examine large-scale datasets. Colossus streamlines the previously labor-intensive data preparation and exploration stages of big data analysis as datasets move through the data pipeline.

Through simple options and the familiar command-line interface, Colossus extends the existing functionalities of Spark and adds new ones to improve early stage data analysis. The most significant original features include the robust file reader that can automatically detect file types as well as classify the data into different groups. The user can further explore different attributes of their data through a data summary as well as visualize those attributes graphically through a browser interface.

**RESULTS**

Colossus shortens the data analytics pipeline, improves early data understanding, and simplifies data preparation. This helps data scientists and data engineers create better data insights faster.

Colossus provides users the ability to analyze and manipulate raw, unclean datasets of various types and from different sources. Through a simple command line interface, users can easily transform and visualize massive data sets without having to write code.

Colossus is more robust than traditional big data tools and is able to process datasets with missing values or illegal values. This allows data scientists to visualize their dataset before they are cleaned.

**IMPACT**

Colossus shortens the data analytics pipeline, improves early data understanding, and simplifies data preparation. This helps data scientists and data engineers create better data insights faster.

Colossus provides users the ability to analyze and manipulate raw, unclean datasets of various types and from different sources. Through a simple command line interface, users can easily transform and visualize massive data sets without having to write code.

Colossus is more robust than traditional big data tools and is able to process datasets with missing values or illegal values. This allows data scientists to visualize their dataset before they are cleaned.