**Data Ingestion**

This project aim to fetch the data from different Stock markets across the world and store the data in NoSql database like Cassandra and HBase.

**This document is not complete yet and will be updated as the project progress.**

1. **Input data file details**

We are fetching NYSE data using Yahoo Financial API and are considering the following fields as of now.

Stock Exchange

Stock Symbol

Trade Date

Open

Previous Close

Day’s low

Day’s high

P/E Ratio

52-weeks low

52-weeks high

Change & percent change

Earnings/Share

Dividend/Share

Volume

Book Value

Bid size

1. **Schema**

Keyspace – Stock Market

Column Family - Stock exchange

RowKey - Stock\_symbol

ColumnKey - Date/Month/Year

Column - Fields

* 1. **Reason for RowKey selection**

According to the financial analysis project problem statement, we will need to query by Stock\_symbol and Timestamp. If we take timestamp as RowKey, then we might end up using a particular set of node more frequently (overloading particular nodes in the cluster) and some nodes may not be used at all. If we consider Stock\_symbol as the RowKey then the load will be possibly balanced.

Another option would be composite key of Stock\_symbol & Year.

Further Analysis is required and this section will be updated once we finalize the complete schema after testing the possibilities. Also the usage of ColumnKey needs more analysis.