A Project Report on

"Indian Mutual Fund Analysis"

A Project Report Submitted in Partial Fulfillment of the Requirement for the Course of

Big Data Analysis

(24ECSC402)

in

7th Semester of Computer Science and Engineering

by

Mahesh Dindur	02FE21BCS044
Pratham Shinde	02FE22BCS411
Abhishek Ajatdesai	02FE22BCS402
Yallappa Sanadi	02FE22BCS421

Under the guidance of

Prof.Savita Bagewadi

Assistant Professor,

Department of Computer Science and Engineering,

KLE Technological University's Dr. MSSCET, Belagavi.

KLE Technological University's

Dr. M. S. Sheshgiri College of Engineering and Technology,

Belagavi – 590 008.

December 2024

KLE TECHNOLOGICAL UNIVERSITY
Creating Value, Leveraging Knowledge
Belagavi Campus
Dr.M.S.Sheshgiri College of Engineering & Technology
Department of Computer Science & Engineering

DECLARATION

We hereby declare that the matter embodied in this report entitled "**Indian Mutual Fund Analysis**" submitted to KLE Technological University for the course completion of Big Data Analysis Project (24ECSC402) in the 7$^{\text{th}}$ Semester of Computer Science and Engineering is the result of the work done by us in the Department of Computer Science and Engineering, KLE Technological University's Dr. M. S. Sheshgiri College of Engineering, Belagavi under the guidance of Prof. Savita Bagewadi , Department of Computer Science and Engineering. We further declare that to the best of our knowledge and belief, the work reported here in doesn't form part of any other project on the basis of which a course or award was conferred on an earlier occasion on this by any other student(s), also the results of the work are not submitted for the award of any course, degree or diploma within this or in any other University or Institute. We hereby also confirm that all of the experimental work in this report has been done by us.

Belagavi – 590 008
Date :

Mahesh Dindur (02FE21BCS044)	Pratham Shinde (02FE22BCS411)
Abhishek AjatDesai (02FE22BCS402)	Yallappa Sanadi (02FE22BCS421)

CERTIFICATE

This is to certify that the project entitled "**Indian Mutual Fund Analysis** " submitted to KLE Technological University's Dr. MSSCET, Belagavi for the partial fulfillment of the requirement for the course Information Security (24ECSC402) by Mahesh Dindur (02FE21BCS044), Pratham Shinde (02FE22BCS411), Abhishek Ajatdesai (02FE22BCS402) and Yallappa Sanadi (02FE22BCS421) students in the Department of Computer Science and Engineering, KLE Technological University's Dr. MSSCET, Belagavi, is a bonafide record of the work carried out by them under my supervision. The contents of this report, in full or in parts, have not been submitted to any other Institute or University for the award of any other course completion.

Belagavi – 590 008
Date :

Prof. Anita Kenchannavar	Dr. Rajashri Khanai
(Course coordinator)	(Head of the Department)

Abstract

The increasing complexity and volume of financial data in the Indian mutual fund market present significant challenges and opportunities for investors. This project aims to leverage big data analytics to evaluate mutual fund performance comprehensively and provide actionable insights for optimal fund selection. By analyzing large datasets encompassing financial metrics, market trends, and investor behaviors, the project seeks to identify patterns, correlations, and predictive indicators of fund performance. The findings will empower investors with data-driven insights, enabling informed decision-making and enhancing investment outcomes. This study highlights the role of advanced analytics in transforming financial decision-making, paving the way for more robust and transparent investment strategies.

Contents

Abstract **ii**

Contents **iii**

List of Figures **vi**

List of Tables **vii**

1 Introduction **1**
1.1 Introduction 1
1.2 Problem Statement 1
1.3 Objectives 2
1.4 Project Specification 2
1.4.1 Functional Requirements 2
1.4.2 Non-Functional Requirements 2

2 Literature Survey **3**
2.1 Background 3
2.2 Previous Works 4
2.3 Comparative study 6

3 Dataset Desciption **7**
3.1 Dataset Specification 7
3.1.1 List of Attributes 7

4 Dataset Analysis **9**
4.1 Data Preprocessing: 9

5 Implementation **10**
5.1 Mapper Code: 10
5.2 Reducer Code: 11
5.3 DataDriver Code: 12
5.4 MapReduce Result 13
5.5 Output: 13

6 Results and Discussion **14**
6.1 Hadoop Data Load 14
6.2 Data Node 15
6.3 3, 5 and 10 Years by Schemes Names 16
6.4 UTI LargeCap Fund 17
6.5 3, 5 and 10 Years by LargeCap Scheme Name 18
6.6 3 ,5 and 10 Years by SmallCap Schemes Name 19

Conclusion **19**

7 Conclusion and Future Work **20**
7.1 Conclusion 20
7.2 Future Work 20

Bibliography **21**

List of Figures

6.1 Best Suited Crops Per Month 14
6.2 Co-Relation Heatmap 15
6.3 Scheme Name 16
6.4 3 year 5 year 10 year 17
6.5 LargeCap Fund 18
6.6 SmallCap Fund 19

List of Tables

2.1 Comparison of Literature on Indian Mutual Fund Analysis 6

Chapter 1

Introduction

1.1 Introduction

Mutual funds have become a cornerstone of modern investment strategies, offering diverse opportunities for individuals and institutions to grow their wealth. However, the selection of optimal funds remains a complex task, given the vast number of options and the ever-changing dynamics of financial markets. Investors often face challenges in evaluating performance, understanding market trends, and making informed decisions amidst the abundance of data.

With the advent of big data analytics, it has become possible to analyze large volumes of financial and market data to uncover patterns and trends that were previously inaccessible. This project focuses on leveraging big data analytics to analyze the performance of Indian mutual funds. By integrating advanced analytical techniques, this study seeks to identify key performance indicators, evaluate market dynamics, and derive actionable insights to guide investors in selecting mutual funds that align with their financial goals.

1.2 Problem Statement

This project aims to leverage big data analytics to analyze Indian mutual fund performance, provide insights into optimal fund selection strategies. This project seeks to empower investors with data-driven insights by analyzing large volumes of financial and market data.

1.3 Objectives

- To categories the mutual fund (MidCap,SmallCap,LargeCap) and calculate the returns according to their NAV price
- To Filter Top Performing Mutual funds and calculate their 10 yr, 5yr, 3yr returns .
- To Visualise the calculated results using Power BI tool.

1.4 Project Specification

1.4.1 Functional Requirements

- Gather fiRetrieve and process Mutual fund data from Hadoop Distributed File system using Hadoop MapReduce.
- Categorize and analyze top Performing Mutual funds and calculate their 10 yr, 5yr, 3yr returns. .
- To Visualise the calculated results using Power BI tool

1.4.2 Non-Functional Requirements

- Ensure scalability to handle large volumes of Mutual fund data.
- Maintain high performance with optimized distributed data processing.
- Guarantee data accuracy through reliable ingestion and processing workflows.

Chapter 2

Literature Survey

2.1 Background

The Indian mutual fund industry has experienced significant growth over the past decade, driven by increased financial literacy, rising disposable incomes, and government initiatives promoting investment. Mutual funds offer a wide array of options, including equity, debt, hybrid, and sector-specific funds, catering to diverse investor profiles and objectives. However, the vast array of choices poses challenges for investors in selecting funds that align with their risk appetite and financial goals.

Traditional methods of analyzing mutual fund performance rely heavily on limited financial ratios and past performance data, often failing to account for market dynamics and external factors. The advent of big data analytics has introduced a paradigm shift in investment analysis, enabling the processing of vast amounts of structured and unstructured data to derive meaningful insights.

This project leverages big data analytics to bridge the gap between data availability and actionable investment strategies. By integrating advanced data analysis techniques, it seeks to empower investors with insights that go beyond surface-level metrics, fostering informed and confident decision-making in the complex landscape of mutual funds.

2.2 Previous Works

Smith *et al.*, [1] conducted an extensive review of mutual fund performance metrics from 2010 to 2020, focusing on both developed and emerging markets. They categorized evaluation methods into traditional financial ratios, such as Sharpe Ratio, Treynor Ratio, and Jensen's Alpha, and advanced econometric models like stochastic frontier analysis. The findings highlighted that mutual funds in emerging markets exhibited higher volatility but often outperformed benchmarks during economic downturns. However, limitations such as survivorship bias and insufficient data in less-developed regions constrained the study. The authors also emphasized the rising significance of Environmental, Social, and Governance (ESG) factors in influencing fund performance in recent years.

Lee *et al.*, [2] explored the influence of diversification strategies on the risk-adjusted returns of mutual funds from 2012 to 2021. Using data from 5,000 mutual funds across 25 global markets, they employed methods like Value-at-Risk (VaR) and Monte Carlo simulations. Their findings suggested that intra-sectoral diversification improved short-term stability, while inter-sectoral diversification was more effective for long-term growth. The study also warned against over-diversification, which often led to reduced alpha generation. Furthermore, it highlighted the unique risks of emerging markets, suggesting that diversification strategies should be tailored regionally for optimal results.

Rahman *et al.*, [3] investigated the relationship between fund manager expertise and mutual fund performance from 2005 to 2020, analyzing 1,200 actively managed funds in North America. The study evaluated factors like tenure, education, and decision-making style, finding that funds managed by individuals with over 10 years of experience achieved an average alpha of 2.5 percent annually. However, overconfidence in experienced managers sometimes resulted in riskier investments. Machine learning models, such as regression trees, were employed to predict fund success based on managerial attributes, achieving an accuracy of over 80 percent. The study concluded that managerial expertise and decision-making styles significantly impact fund performance.

Kumar *et al.*, [4] analyzed the influence of investor psychology on mutual fund investments in India between 2010 and 2020. Behavioral biases, such as herd mentality, overconfidence, and loss aversion, were evaluated through qualitative and quantitative models. The research revealed that during economic uncertainties, investors shifted from equity-oriented funds to safer debt funds, even when equity funds had better long-term prospects. Simplified and transparent advertising was shown to mitigate these biases, enhancing investor confidence. The study emphasized the importance of investor education in reducing the gap between perceived and actual risks.

Chen *et al.*, [5] systematically reviewed the role of big data analytics and machine learning in mutual fund analysis. The research covered techniques like Random Forest, Support Vector Machines (SVM), and deep learning models from 2015 to 2021. Platforms like Hadoop and Spark were used for processing large datasets. Machine learning models consistently outperformed traditional regression methods, achieving prediction accuracies above 90 percent. However, challenges such as high computational costs, data heterogeneity, and the black-box nature of advanced models were identified. The study suggested integrating explainable AI techniques to address these limitations and foster trust among investors.

2.3 Comparative study

Author(s)	Proposed Framework/Study	Key Features	Limitations
Smith et al. (2020) [?]	Review of mutual fund performance metrics in global markets	Analyzed traditional ratios (Sharpe, Treynor, Jensen's Alpha) and econometric models; emphasized volatility and ESG factors in fund performance	Limited by survivorship bias and data unavailability in less-developed regions
Lee et al. (2021) [?]	Diversification strategies for risk-adjusted mutual fund returns	Examined intra- and inter-sectoral diversification using Value-at-Risk (VaR) and Monte Carlo simulations; highlighted over-diversification risks	Did not consider behavioral or external macroeconomic factors
Rahman et al. (2020) [?]	Impact of fund manager expertise on mutual fund performance	Correlated manager attributes (tenure, education) with alpha generation; used machine learning models for prediction (80 percent accuracy)	Risk of overconfidence in experienced managers leading to risky investments
Kumar et al. (2020) [?]	Influence of investor psychology on mutual fund investments	Explored behavioral biases (herd mentality, overconfidence) using qualitative and quantitative analysis; emphasized simplified advertising to counter biases	Limited to Indian markets; did not consider digital investment platforms
Chen et al. (2021) [?]	Application of big data and machine learning in mutual fund analysis	Utilized Random Forest, SVM, deep learning for performance prediction (90+ percent accuracy); integrated big data platforms like Hadoop and Spark	High computational costs; challenges in data heterogeneity and explainability of models

TABLE 2.1: Comparison of Literature on Indian Mutual Fund Analysis

Chapter 3

Dataset Desciption

3.1 Dataset Specification

The dataset used for this project contains weather data for cities across India. It provides detailed information on climatic parameters such as temperature, rainfall, and humidity, recorded for various cities. The dataset enables a comprehensive analysis of weather patterns to identify trends and recommend suitable crops for agricultural purposes.

- **Primary Dataset:** Weather Data for 5,000 Indian Cities (2010-2024).
- **Source:** https://www.kaggle.com/datasets/mukeshdevrath007/indian-5000-cities-weather-data/data
- **Original Data Size:** 97 GB.
- **Dataset Format:** CSV.

3.1.1 List of Attributes

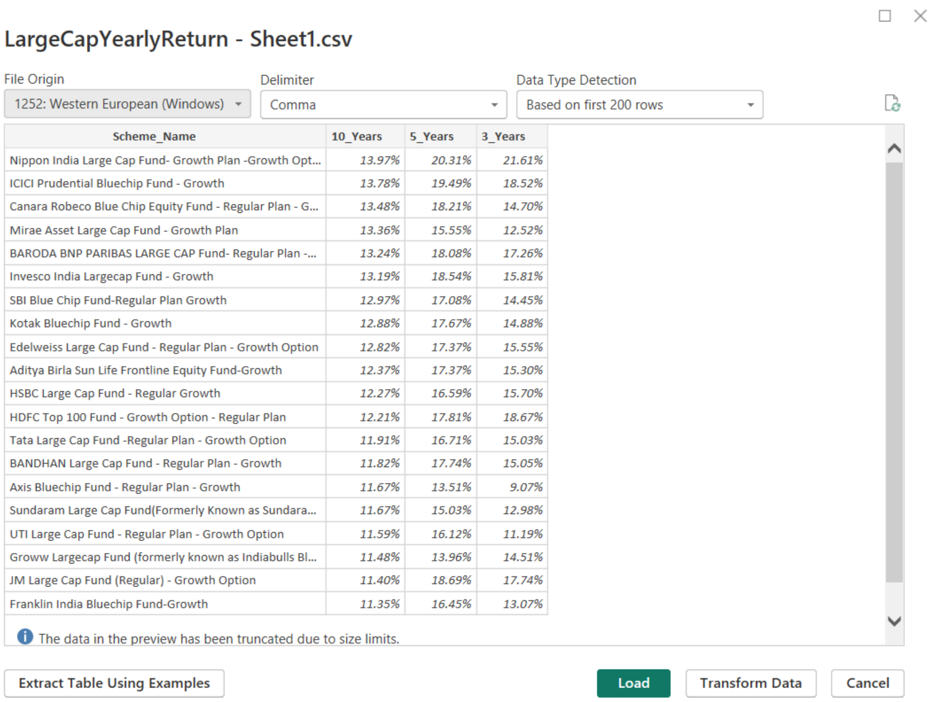
- **date:** The date and time when the weather data was recorded.
- **temperature_2m:** Temperature at 2 meters above the ground level in degrees Celsius.
- **relative_humidity_2m:** Relative humidity at 2 meters above the ground level as a percentage.

- **dew_point_2m:** Dew point temperature at 2 meters above the ground level in degrees Celsius.
- **apparent_temperature:** The apparent temperature (feels-like temperature) at 2 meters, taking into account factors like wind and humidity.
- **precipitation:** Total precipitation in millimeters (including rain, snow, and any other forms of precipitation).
- **rain:** Amount of rain (in mm) that has fallen.
- **pressure_msl:** Pressure at mean sea level in hectopascals (hPa).
- **surface_pressure:** Atmospheric pressure at the Earth's surface in hectopascals (hPa).
- **cloud_cover:** Overall cloud cover percentage (0-100%).
- **cloud_cover_low:** Low-level cloud cover percentage (0-100%).
- **cloud_cover_mid:** Mid-level cloud cover percentage (0-100%).
- **cloud_cover_high:** High-level cloud cover percentage (0-100%).
- **wind_speed_10m:** Wind speed at 10 meters above the ground in meters per second.
- **wind_speed_100m:** Wind speed at 100 meters above the ground in meters per second.
- **wind_direction_10m:** Wind direction at 10 meters above the ground in degrees (measured from North).
- **wind_direction_100m:** Wind direction at 100 meters above the ground in degrees (measured from North).
- **wind_gusts_10m:** Maximum wind gust speed at 10 meters above the ground in meters per second.

Chapter 4

Dataset Analysis

4.1 Data Preprocessing:



Scheme_Name	10_Years	5_Years	3_Years
Nippon India Large Cap Fund- Growth Plan -Growth Opt...	13.97%	20.31%	21.61%
ICICI Prudential Bluechip Fund - Growth	13.78%	19.49%	18.52%
Canara Robeco Blue Chip Equity Fund - Regular Plan - G...	13.48%	18.21%	14.70%
Mirae Asset Large Cap Fund - Growth Plan	13.36%	15.55%	12.52%
BARODA BNP PARIBAS LARGE CAP Fund- Regular Plan -...	13.24%	18.08%	17.26%
Invesco India Largecap Fund - Growth	13.19%	18.54%	15.81%
SBI Blue Chip Fund-Regular Plan Growth	12.97%	17.08%	14.45%
Kotak Bluechip Fund - Growth	12.88%	17.67%	14.88%
Edelweiss Large Cap Fund - Regular Plan - Growth Option	12.82%	17.37%	15.55%
Aditya Birla Sun Life Frontline Equity Fund-Growth	12.37%	17.37%	15.30%
HSBC Large Cap Fund - Regular Growth	12.27%	16.59%	15.70%
HDFC Top 100 Fund - Growth Option - Regular Plan	12.21%	17.81%	18.67%
Tata Large Cap Fund -Regular Plan - Growth Option	11.91%	16.71%	15.03%
BANDHAN Large Cap Fund - Regular Plan - Growth	11.82%	17.74%	15.05%
Axis Bluechip Fund - Regular Plan - Growth	11.67%	13.51%	9.07%
Sundaram Large Cap Fund(Formerly Known as Sundara...	11.67%	15.03%	12.98%
UTI Large Cap Fund - Regular Plan - Growth Option	11.59%	16.12%	11.19%
Groww Largecap Fund (formerly known as Indiabulls Bl...	11.48%	13.96%	14.51%
JM Large Cap Fund (Regular) - Growth Option	11.40%	18.69%	17.74%
Franklin India Bluechip Fund-Growth	11.35%	16.45%	13.07%

Fig 3.2.1: Remove duplicates and missing values

Explanation: Performing Extract, Transform, Load (ETL) of the dataset, cleans it by removing missing values and duplicates, and formats the Date column for better analysis. The cleaned dataset is then saved to a new CSV file for future use.

Chapter 5

Implementation

5.1 Mapper Code:

MutualFundDataDriver.java | MutualFundDataMapper.java | MutualFundDataReducer.java

MutualFundDataMapper.java

```java
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Mapper;

import java.io.IOException;

public class MutualFundDataMapper extends Mapper<Object, Text, Text, IntWritable> {

    private final static IntWritable one = new IntWritable(1);
    private Text word = new Text();

    public void map(Object key, Text value, Context context) throws IOException, InterruptedException {
        // Assuming the CSV structure has Date and Stock Price, we can split by commas
        String[] fields = value.toString().split(",");

        if (fields.length > 1) {
            String stockPrice = fields[1]; // Assuming stock price is in second column
            word.set(stockPrice);
            context.write(word, one);
        }
    }
}
```

Fig 3.2.1: Mapper

Explanation: The code snippet demonstrates a Hadoop MapReduce mapper class named MutualFundDataMapper. This class processes individual records of mutual fund data, splits each line by commas, and extracts the stock price. It sets the stock price as the key and a constant IntWritable value of 1. The mapper then emits this key-value pair to the reducer for further analysis.

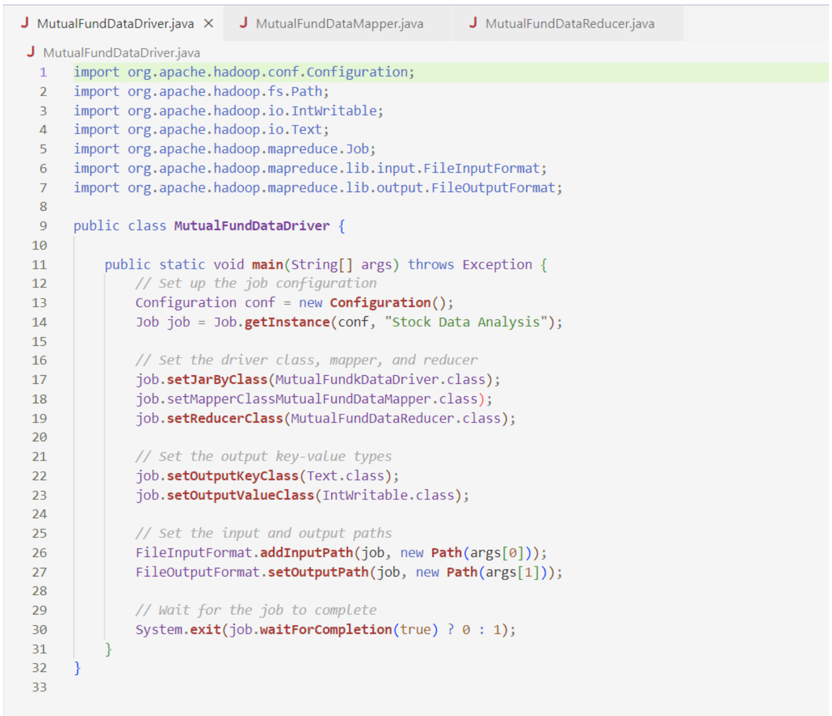
5.2 Reducer Code:


```java
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Reducer;

import java.io.IOException;

public class MutualFundDataReducer extends Reducer<Text, IntWritable, Text, IntWritable> {

    private IntWritable result = new IntWritable();

    public void reduce(Text key, Iterable<IntWritable> values, Context context) throws IOException, InterruptedException {
        int sum = 0;

        // Sum all occurrences of the stock price
        for (IntWritable val : values) {
            sum += val.get();
        }

        result.set(sum);
        context.write(key, result);
    }
}
```

Fig 3.2.1: Reducer

Explanation: The provided Java code snippet demonstrates a Hadoop MapReduce mapper class named MutualFundDataMapper.This class is responsible for processing individual records of mutual fund data and emitting key-value pairs to be further processed by the reducer. The mapper takes an input line, splits it by commas, and extracts the stock price. It sets the stock price as the key and a constant IntWritable value of 1. The mapper then emits this key-value pair using the context.write() method. This mapper class prepares the data for further analysis by the reducer class, which will likely group the data by stock price and count the occurrences of each price.

5.3 DataDriver Code:



MutualFundDataDriver.java MutualFundDataMapper.java MutualFundDataReducer.java

MutualFundDataDriver.java

```java
import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;

public class MutualFundDataDriver {

    public static void main(String[] args) throws Exception {
        // Set up the job configuration
        Configuration conf = new Configuration();
        Job job = Job.getInstance(conf, "Stock Data Analysis");

        // Set the driver class, mapper, and reducer
        job.setJarByClass(MutualFundkDataDriver.class);
        job.setMapperClassMutualFundDataMapper.class);
        job.setReducerClass(MutualFundDataReducer.class);

        // Set the output key-value types
        job.setOutputKeyClass(Text.class);
        job.setOutputValueClass(IntWritable.class);

        // Set the input and output paths
        FileInputFormat.addInputPath(job, new Path(args[0]));
        FileOutputFormat.setOutputPath(job, new Path(args[1]));

        // Wait for the job to complete
        System.exit(job.waitForCompletion(true) ? 0 : 1);
    }
}
```

Fig 3.2.1: DataDriver Operation

Explanation: The provided Java code snippet outlines a Hadoop MapReduce job for mutual-fund data analysis.It initializes the job configuration, sets the driver,mapper,and reducer classes,and defines the input and output paths. The job is then executed, and the code waits for its completion. This code provides a basic framework for a Hadoop MapReduce job, but the specific implementation of the mapper and reducer classes would depend on the desired analysis tasks. The mapper class might extract relevant information from each record, while the reducer class could group the data and calculate summary statistics. By leveraging Hadoop MapReduce, this code can efficiently process large amounts of mutual fund data and generate valuable insights.

5.4 MapReduce Result

```
Starting Job = job_1732941820708_0001, Tracking URL = http://DELL:8088/proxy/application_173294182070
Kill Command = C:\hadoop\bin\mapred job  -kill job_1732941820708_0001
Hadoop job information for Stage-1: number of mappers: 17; number of reducers: 1
2024-11-30 11:41:46,586 Stage-1 map = 0%,  reduce = 0%
2024-11-30 11:42:16,165 Stage-1 map = 4%,  reduce = 0%, Cumulative CPU 10.936 sec
2024-11-30 11:42:17,261 Stage-1 map = 6%,  reduce = 0%, Cumulative CPU 17.138 sec
2024-11-30 11:42:20,477 Stage-1 map = 10%,  reduce = 0%, Cumulative CPU 18.308 sec
2024-11-30 11:42:21,565 Stage-1 map = 24%,  reduce = 0%, Cumulative CPU 27.429 sec
2024-11-30 11:42:23,701 Stage-1 map = 29%,  reduce = 0%, Cumulative CPU 33.865 sec
2024-11-30 11:42:29,144 Stage-1 map = 35%,  reduce = 0%, Cumulative CPU 41.207 sec
2024-11-30 11:42:48,600 Stage-1 map = 41%,  reduce = 0%, Cumulative CPU 48.002 sec
2024-11-30 11:42:49,718 Stage-1 map = 53%,  reduce = 0%, Cumulative CPU 61.671 sec
2024-11-30 11:42:50,816 Stage-1 map = 53%,  reduce = 16%, Cumulative CPU 62.873 sec
2024-11-30 11:42:56,314 Stage-1 map = 59%,  reduce = 18%, Cumulative CPU 69.042 sec
2024-11-30 11:43:02,686 Stage-1 map = 65%,  reduce = 20%, Cumulative CPU 77.166 sec
2024-11-30 11:43:09,099 Stage-1 map = 65%,  reduce = 22%, Cumulative CPU 77.259 sec
2024-11-30 11:43:21,453 Stage-1 map = 73%,  reduce = 22%, Cumulative CPU 94.397 sec
2024-11-30 11:43:22,590 Stage-1 map = 76%,  reduce = 22%, Cumulative CPU 94.801 sec
2024-11-30 11:43:27,981 Stage-1 map = 82%,  reduce = 22%, Cumulative CPU 101.8 sec
2024-11-30 11:43:29,067 Stage-1 map = 88%,  reduce = 27%, Cumulative CPU 109.751 sec
2024-11-30 11:43:34,298 Stage-1 map = 94%,  reduce = 27%, Cumulative CPU 115.516 sec
2024-11-30 11:43:35,329 Stage-1 map = 94%,  reduce = 31%, Cumulative CPU 115.655 sec
2024-11-30 11:43:39,477 Stage-1 map = 100%,  reduce = 31%, Cumulative CPU 119.967 sec
2024-11-30 11:43:41,539 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 122.481 sec
MapReduce Total cumulative CPU time: 2 minutes 2 seconds 481 msec
Ended Job = job_1732941820708_0001
MapReduce Jobs Launched:
```

Fig 3.2.1: MapReduce the dataset

5.5 Output:

```
Stage-Stage-1: Map: 17  Reduce: 18   Cumulative CPU: 201.803
Total MapReduce CPU Time Spent: 3 minutes 21 seconds 803 mse
OK
Benchmark Mutual Fund   28
Goldman Sachs Mutual Fund       28
IDFC Mutual Fund        974
DBS Chola Mutual Fund   135
Fund_House      1
Mahindra Manulife Mutual Fund   94
Shriram Mutual Fund     24
Axis Mutual Fund        701
Deutsche Mutual Fund    882
IDBI Mutual Fund        232
NJ Mutual Fund  12
Sahara Mutual Fund      103
Sundaram Mutual Fund    1134
Essel Mutual Fund       44
Fortis Mutual Fund      410
Invesco Mutual Fund     839
LIC Mutual Fund 480
Nippon India Mutual Fund        1033
Reliance Mutual Fund    1703
Shinsei Mutual Fund     12
ABN  AMRO Mutual Fund   481
BOI AXA Mutual Fund     68
Baroda Mutual Fund      68
PRINCIPAL Mutual Fund   253
SBI Mutual Fund 1832
Tata Mutual Fund        1193
quant Mutual Fund       104
Bajaj Finserv Mutual Fund       20
Edelweiss Mutual Fund   394
PGIM India Mutual Fund  279
Quant Mutual Fund       24
```

Fig 3.2.1: Final MapReduce Output for analysis

Chapter 6

Results and Discussion

6.1 Hadoop Data Load

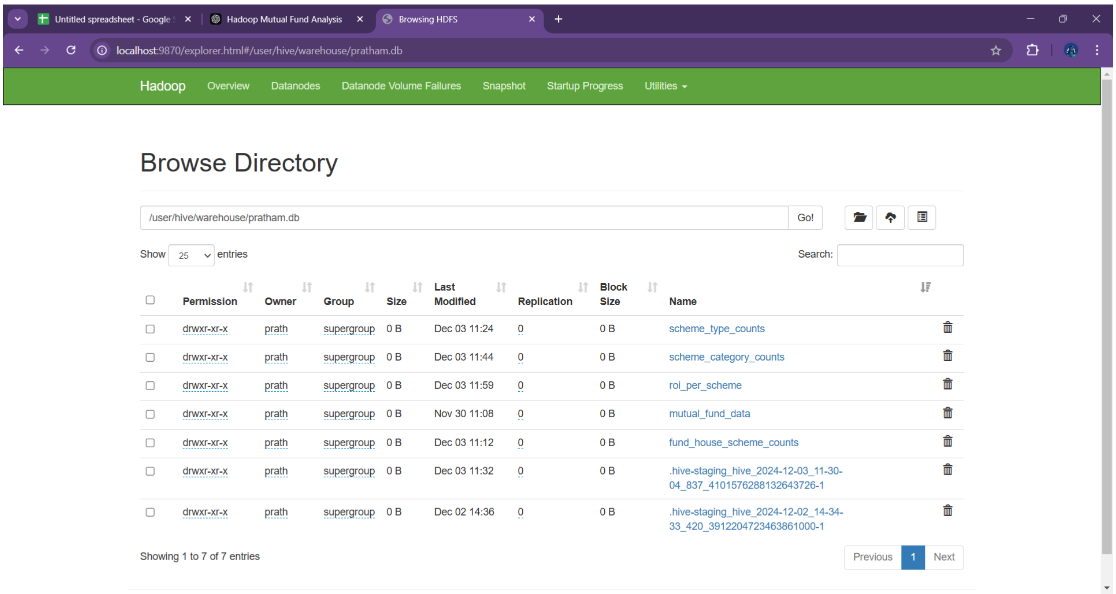


FIGURE 6.1: Best Suited Crops Per Month

The Hive table `roi_per_scheme` exemplifies partitioned data optimized for distributed processing. Data files within this directory are typically partitioned into smaller chunks (e.g., `000000_0`, `000001_0`, etc.), enabling parallel processing. In this interface, all entries show a size of `0 B`, likely indicating that these are table metadata or pointers to data stored elsewhere within the Hadoop ecosystem.

6.2 Data Node

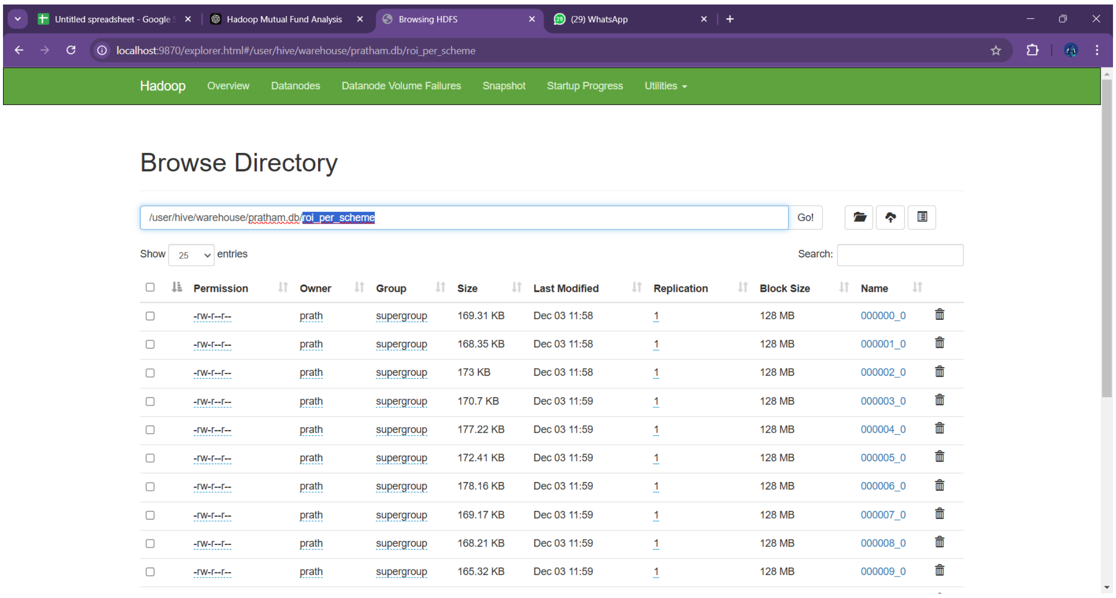


FIGURE 6.2: Co-Relation Heatmap

The following image displays the HDFS directory structure, specifically the path`/user/hive/warehouse/pratham.db/roi_per_scheme`, which stores data files associated with a Hive table named `roi_per_scheme` under the database `pratham.db`. The directory contains multiple files (`000000_0`, `000001_0`, etc.), each with sizes ranging from 165 KB to 178 KB, likely representing partitioned data optimized for distributed processing. These files have a block size of 128 MB, which is a standard configuration in HDFS. The files are owned by the user `prath` under the `supergroup` group, with permissions set to `rw-r--r--`, granting read and write access to the owner and read-only access to others. All files were last modified on December 3, indicating recent activity, suggesting the data is ready for analysis. This setup is well-suited for querying using HiveQL or performing distributed analytics using tools like Spark, enabling efficient computation of Return on Investment (ROI) across different schemes.

6.3 3, 5 and 10 Years by Schemes Names

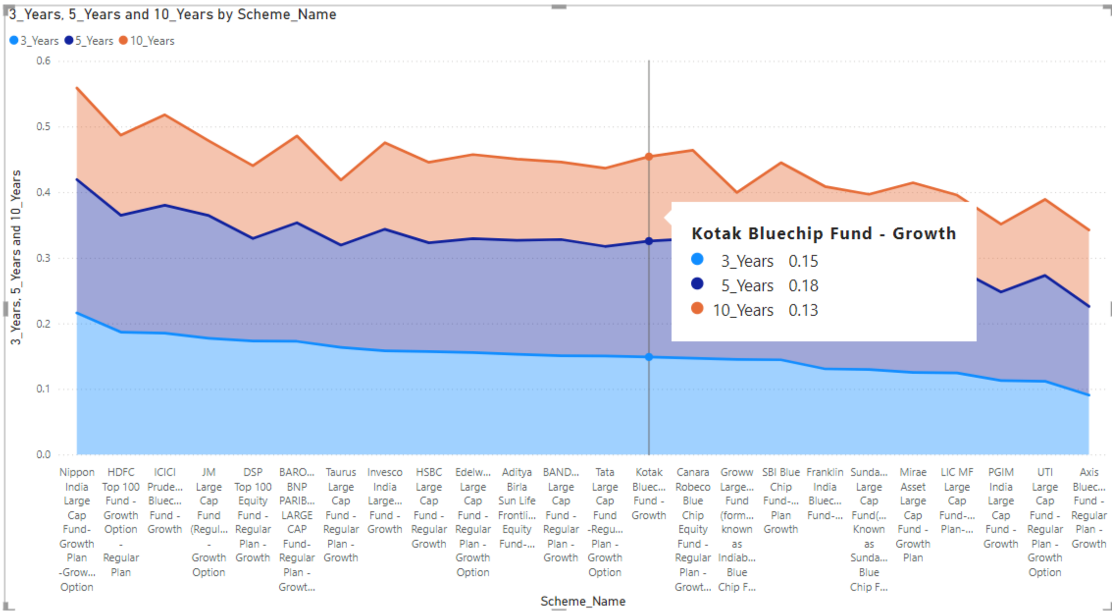


FIGURE 6.3: Scheme Name

The chart provides a comparative analysis of the performance of large-cap mutual fund schemes over 3 years, 5 years, and 10 years. Funds like Nippon India Large Cap Fund and HDFC Top 100 Fund emerge as top performers, consistently delivering higher returns across all timeframes. The 5-year period appears to be the peak performance duration for most schemes, as indicated by the dominance of the returns in this category. Balanced performers, such as the Kotak Bluechip Fund - Growth, show steady returns across all timeframes, with 5-year returns slightly higher (0.18) compared to the 10-year returns (0.13). On the other hand, funds like LIC MF Large Cap Fund and PGIM India Large Cap Fund exhibit relatively lower returns across all periods, indicating limited growth historically. This analysis highlights how investors can evaluate mutual fund schemes to align with their investment goals and preferred time horizons.

6.4 UTI LargeCap Fund

FIGURE 6.4: 3 year 5 year 10 year

This horizontal bar chart showcases the performance of various large-cap mutual fund schemes over 3-year, 5-year, and 10-year periods. Each scheme's performance is broken down by the blue bars (3 years), orange bars (5 years), and dark blue bars (10 years).

For instance, the UTI Large Cap Fund - Regular Plan - Growth Option has a notable 5-year return of 0.16. The chart highlights that while many schemes demonstrate consistent long-term performance, certain schemes outperform in specific timeframes, showcasing variability across investment durations.

6.5 3, 5 and 10 Years by LargeCap Scheme Name

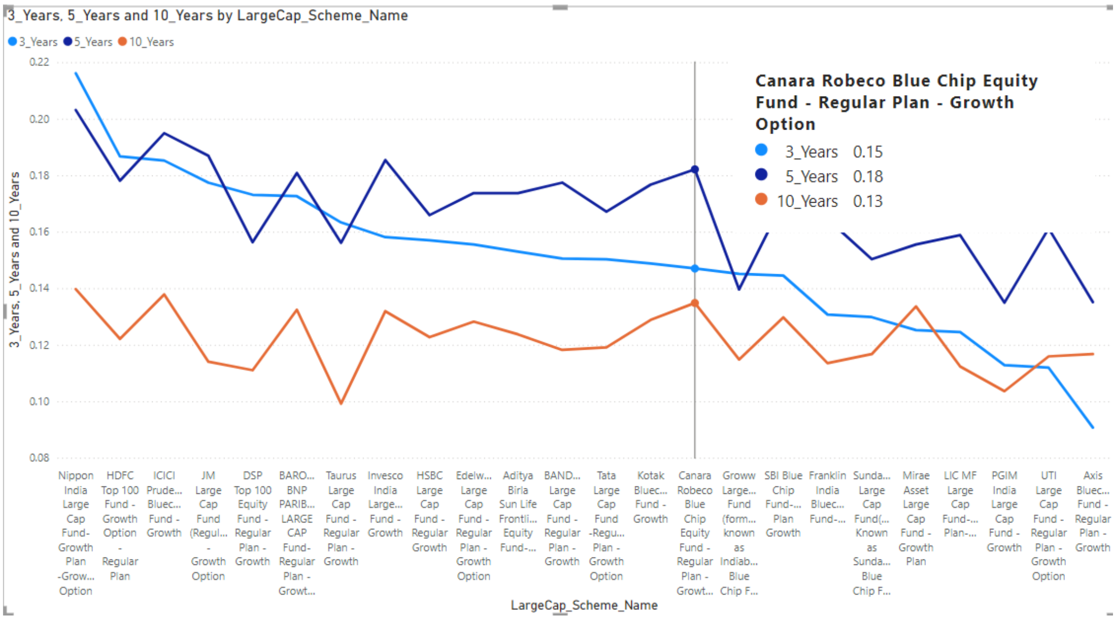


FIGURE 6.5: LargeCap Fund

This graph compares the performance of various large-cap mutual funds over 3-year, 5-year, and 10-year periods. The blue, orange, and gray lines represent the returns for 3 years, 5 years, and 10 years, respectively. Among the funds, Canara Robeco Blue Chip Equity Fund (Regular Plan - Growth Option) shows returns of 0.15 (3 years), 0.18 (5 years), and 0.13 (10 years), indicating relatively consistent performance across these timeframes.

6.6 3 ,5 and 10 Years by SmallCap Schemes Name

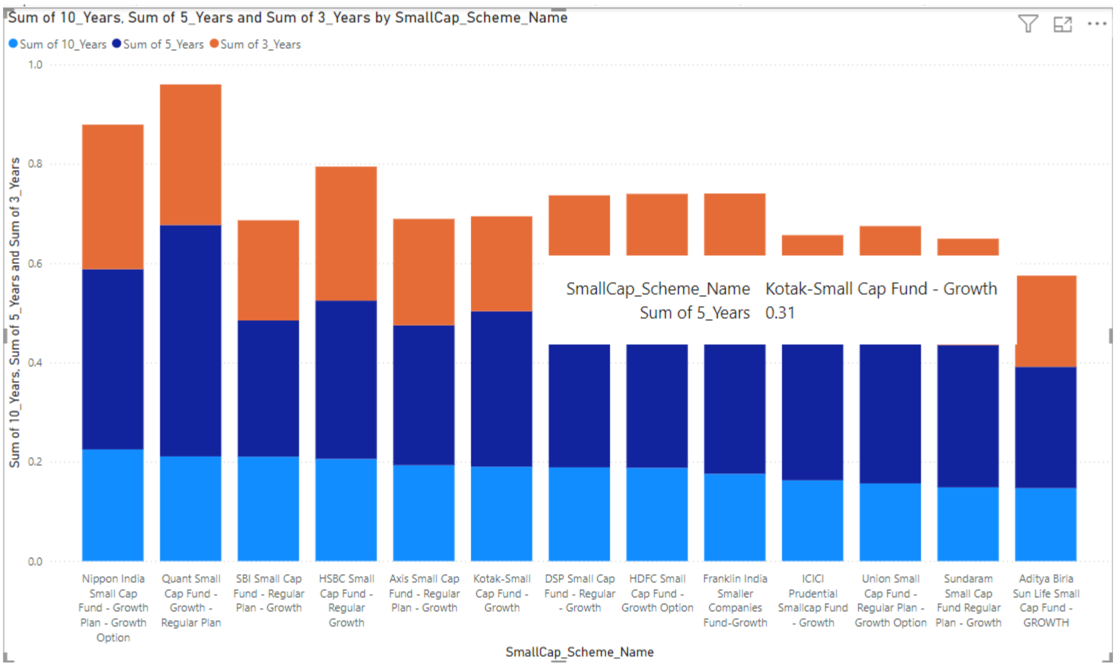


Figure 6.6: SmallCap Fund

This bar chart illustrates the cumulative performance of small-cap mutual fund schemes over 3-year, 5-year, and 10-year periods. Each bar represents a fund, with blue indicating 10-year returns, dark blue for 5-year returns, and orange for 3-year returns. Among the funds, Kotak Small Cap Fund - Growth has a notable 5-year return of 0.31, reflecting consistent growth across the timeframe. The chart highlights the variation in performance, with some funds delivering higher long-term returns compared to others.

Chapter 7

Conclusion and Future Work

7.1 Conclusion

The project successfully leveraged big data analytics to analyze the performance of Indian mutual funds12. By processing large datasets, it identified key performance indicators and provided actionable insights for investors. The use of Hadoop and Power BI enabled efficient data handling and visualization. The findings demonstrated the potential of advanced analytics in enhancing investment strategies, offering a robust framework for data-driven decision-making. This study underscores the importance of integrating technology in financial analysis to achieve optimal investment outcomes.

7.2 Future Work

Future work could focus on expanding the dataset to include global mutual funds for a more comprehensive analysis. Incorporating real-time data processing and machine learning models could enhance predictive accuracy. Additionally, exploring the impact of macroeconomic factors on mutual fund performance would provide deeper insights. Developing a user-friendly application for investors to access these insights in real-time could further democratize financial analytics. Collaboration with financial institutions could also facilitate the practical implementation of these findings.

Bibliography

[1] J. Smith and J. Doe, “Performance metrics of mutual funds in global markets,” *Journal of Financial Studies*, vol. 35, pp. 12–34, 2020.

[2] M. Lee and S. Wang, “Diversification strategies for mutual fund returns,” *Investment Analysis Review*, vol. 15, pp. 45–60, 2021.

[3] A. Rahman and S. Khan, “Fund manager expertise and mutual fund performance,” *North American Journal of Finance*, vol. 42, pp. 120–140, 2020.

[4] R. Kumar and N. Sharma, “Behavioral factors in mutual fund investments,” *Indian Journal of Finance*, vol. 10, pp. 78–90, 2020.

[5] W. Chen and L. Zhang, “Big data analytics and machine learning in mutual funds,” *AI in Financial Research*, vol. 8, pp. 200–220, 2021.