**MSS iThink :**

iThink is a software which is used to analyze data on the basis of attributes available in the data set.

One can analyse the data and perform number of queries on the data set.

This software provides two types of analysis methods:

1. Theoretical Analysis.

2. Graphical Analysis.

**Technologies used:**

1. Python(2.7): A programming language that lets you work more quickly and integrate your systems more effectively.

You can learn to use Python and see almost immediate gains in productivity and lower maintenance costs.

Launched on 17 December 2016, 5 months ago

1. Apache Spark(2.1.1): A fast and general engine for large-scale data processing,

It provides programmers with an [application programming interface](https://en.wikipedia.org/wiki/Application_programming_interface) centered on a [data structure](https://en.wikipedia.org/wiki/Data_structure) called the resilient distributed dataset (RDD), a read-only [multiset](https://en.wikipedia.org/wiki/Multiset) of data items distributed over a cluster of machines, that is maintained in a [fault-tolerant](https://en.wikipedia.org/wiki/Fault-tolerant_computing) way.

Launched on May 2, 2017; 30 days ago

**Installation Steps:**

1. Python: Run the setup file and install the .exe file.
2. Apache Spark:
   1. Download the latest version of Spark by visiting the following link.

[: https://spark.apache.org/downloads.html](:%20https://spark.apache.org/downloads.html )

* 1. Moving Spark software files:
     + Copy the installed files in to C: drive with a folder named Spark.

1. Hadoop :
   1. Download Winutils.exe from the following link: <https://github.com/steveloughran/winutils/tree/master/hadoop-2.6.0/bin>
   2. Create a folder named winutils in C: Drive.
   3. Move the Winutils.exe file into the folder created in the above step.

**Setting the path Variables:**

1. Python :
   * + Right Click on My Computer > Go to System Properties >

Go to Advanced System Settings > Go to Advanced Tab >

Select Environment Variables > Add Variable Name : Path and Variable Value : C:\Python27;

1. Apache Spark :
   * + Right Click on My Computer > Go to System Properties >

Go to Advanced System Settings > Go to Advanced Tab >

Select Environment Variables > Add Variable Name: SPARK\_HOME and Variable Value : C:\Spark;

3) Hadoop:

* + - Right Click on My Computer > Go to System Properties >

Go to Advanced System Settings > Go to Advanced Tab >

Select Environment Variables > Add Variable Name: HADOOP\_HOME and Variable Value : C:\winutils

**Things to Verify before going further :**

1. Check whether Python is installed properly.
   1. Go to the Command Prompt
   2. Type in there : “ python “ and hit enter.
   3. If the installation went fine, it opens the Python Shell and should return this.

Python 2.7.13 (v2.7.13:a06454b1afa1, Dec 17 2016, 20:53:40) [MSC v.1500 64 bit (AMD64)] on win32

Type "help", "copyright", "credits" or "license" for more information.

>>>

* 1. If this doesn’t happen, Please check the Environment Variables and setting the path for them.

1. Check whether the Spark is installed properly.
   1. Go to the Spark folder.
   2. Open bin folder.
   3. Run Command prompt from there.
   4. Type “ Pyspark “ in there.
   5. If the installation went fine, it opens the Python Shell and should return this.

Spark assembly has been built with Hive, including Datanucleus jars on classpath

Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties

15/06/04 15:25:22 INFO SecurityManager: Changing view acls to: hadoop

15/06/04 15:25:22 INFO SecurityManager: Changing modify acls to: hadoop

15/06/04 15:25:22 INFO SecurityManager: SecurityManager: authentication disabled;

ui acls disabled; users with view permissions: Set(hadoop); users with modify permissions: Set(hadoop)

15/06/04 15:25:22 INFO HttpServer: Starting HTTP Server

15/06/04 15:25:23 INFO Utils: Successfully started service 'HTTP class server' on port 43292.

Welcome to

\_\_\_\_ \_\_

/ \_\_/\_\_ \_\_\_ \_\_\_\_\_/ /\_\_

\_\ \/ \_ \/ \_ `/ \_\_/ '\_/

/\_\_\_/ .\_\_/\\_,\_/\_/ /\_/\\_\ version 1.4.0

/\_/

Using Scala version 2.10.4 (Java HotSpot(TM) 64-Bit Server VM, Java 1.7.0\_71)

Type in expressions to have them evaluated.

Spark context available as sc

**Things done:**

1. The main aim of this project is Data Analysis
   * + We can take any data sets,for example : the details of an online shopping site which might include : Order Id’s, Order Details, Customer’s Details,etc.
     + (Or) We can perform analysis on any data set available in (.CSV) Comma Seperated Values.
     + We can perform any Query on the attributes available in the data set, for example : Number Of orders placed from a particular region , Analysis Based on the shipment mode selected by customers,etc.
     + We can even get to know what all possible queries can be performed by selecting an attribute.
     + We can perform queries based on the selection of attributes and type of query.

**Performing Queries:**

1. The queries take the basic syntax of Structured Query Language (MySQL).
2. With the flexibility of MySQL, we can perform any query based on the choice of attributes.
   * + Syntax : select \* from Table\_name where record ='Attribute'

**Plotting Graphs for Better Understanding:**

1. The queries we performed can be very easy to understand, so we made some graphs and pie charts based on the results of the queries.

These can be plotted using Python’s Module named: pygal

Front-End :

* Front end is designed using the flask module of python

[**Flask**](http://flask.pocoo.org/) is a web framework. This means flask provides you with tools, libraries and technologies that allow you to build a web application. This web application can be some web pages, a blog, a wiki or go as big as a web-based calendar application or a commercial website.

In simple, it’s nothing but the integration of interactive UI for python