Project

Data As a Service

Team id: 27

Project number:17

Team Members	Mentor
Deepak Upreti (201405533)	Vishrut Mehta
Gunjit Bansal (201405568)	
Spurthi Tallam (201301241)	
Pradeep Kumar Anumala (201350843)	

Project Goal:

The goal of this project is to build data as a service platform (like data.gov.in) in which data is made available to customers over a network.

It is about storing data that should be highly available, reliable and performant. The user need not worry about the data storage

Approach:

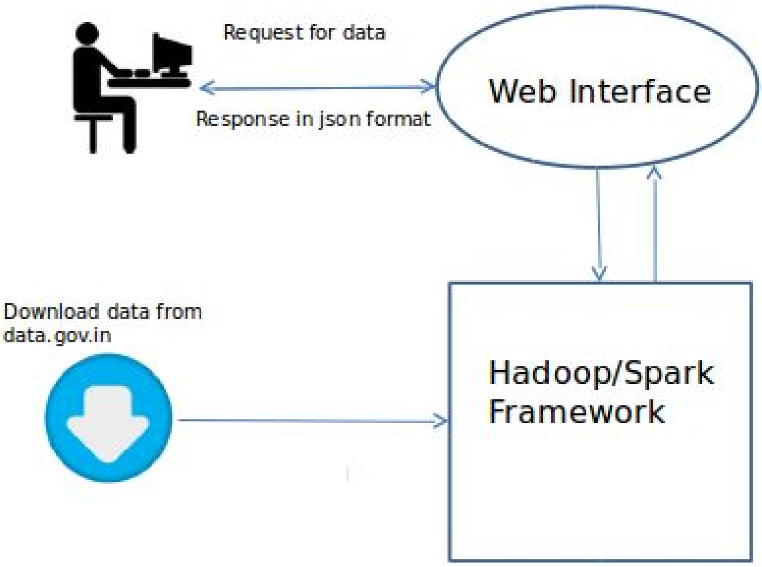
This platform will provide data to the users when requested. Users request to download and upload the data through UI built using web2py.

Availability, reliability are some of the key points that the platform should take care of. In order to handle/process large data, we are using Hadoop HDFS and Spark framework.

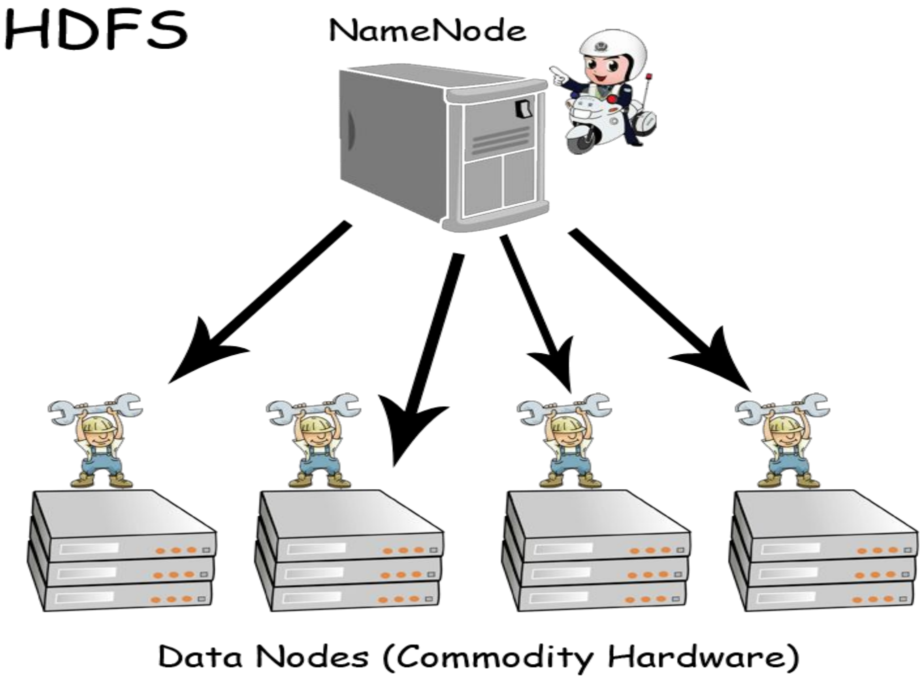
Requirements:

- ➔ Build up a SQL database to store data and information as the query can be specific based on the attributes or fields the client is interested in. Also the database should favor easy updation.
- ➔ Provide the user a decent interface to interact with our platform.
- ➔ The platform should be reliable and able to respond quickly
- ➔ Client should be able to trigger the queries based on the fields or attributes he is interested in, also he can request for full document.
- ➔ Client will be given access to add any document to the cluster.
- ➔ User can however add and download data from HDFS cluster but they won’t be provided rights to modify any existing data on cluster.

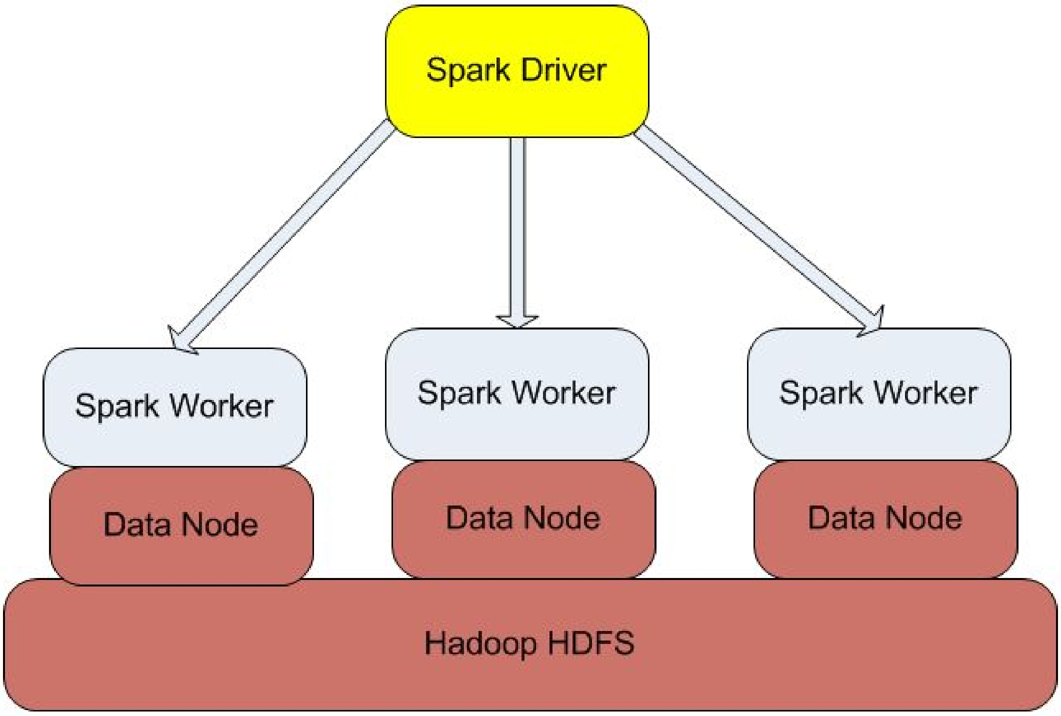
High Level Design:



Overall Approach



Apache Spark Layout



:

HDFS Architecture

Implementation:

- ➔ Installed hadoop (2.6.0),Spark(1.4.0),web2py,CherryPy
- ➔ Set the pythonpath as shown below in bashrc

```
export SPARK_HOME=/usr/local/spark
export PYTHONPATH=$SPARK_HOME/python:$SPARK_HOME/python/build:$PYTHONPATH
```

➔ Web2py role:

Web2py acts as an application server which provides a UI to the user, takes the requests ,forwards the requests to the corresponding server and provides the results to the user.

- ◆ Maintains SQLite database which stores the information like filename, category, location, data types of the fields.

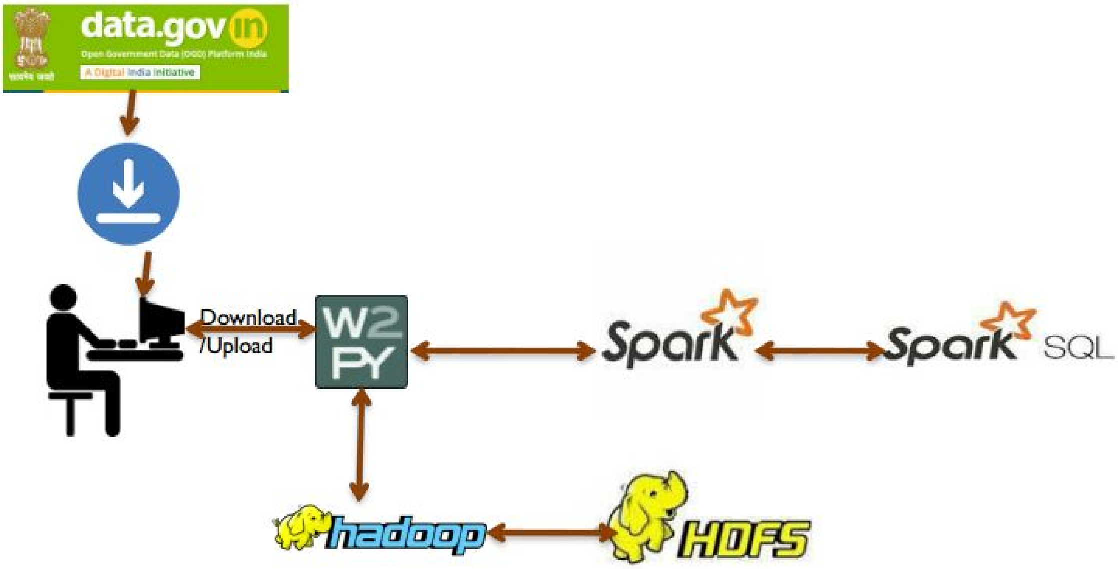
➔ HDFS Cluster role:

- ◆ Whenever user uploads the file, the file goes into the hdfs cluster.
- ◆ HDFS internally takes care of the reliability and availability

➔ Spark role:

- ◆ Spark is responsible for providing the data to the user.
- ◆ When a user tries to download the data, Spark fetches the data from hdfs cluster, creates an RDD, runs the Query on RDD and returns the data.

Process Flow:


data.gov.in
Open Government Data (OGD) Platform India
A Digital India Initiative
Download
/Upload
W2
PY
Spark
Spark SQL
hadoop
HDFS

Upload Data

- The sample data downloaded from data.gov.in is uploaded by the user through User Interface built using web2py.
- Web2py maintains a table that will store all the information about the files like filename, category, location, data types of the fields which will be updated upon successful upload of file.
- The first line of the input file should contain the header (field names). This is used as schema by the Spark.
- The web2py server redirects the request to hadoop hdfs which puts the file into hdfs cluster.
- User must be logged in to upload the data.

Download Data

- The data can be downloaded as an entire file or the user can give filters to fetch the data of his/her interest.
- User need not be logged in to fetch the data.
- User requests the data through web2py UI. Web2py bundles the parameters,creates a query string and forwards it to the spark server.
- The sparkserver interacts with hdfs and fetches the file.
- Spark Server fetches sql context and creates an RDD. An SQL query is run on the RDD to fetch the desired data.
- The output data is returned in the json format.

Results :

- Successfully uploaded file
- Successfully downloaded complete file
- Successfully downloaded file with filters.