

STUDYING THE INDUSTRIAL PATTERN USING HADOOP

MAJOR PROJECT REPORT

*Submitted in partial fulfillment of the requirement for the award of the degree
of*

BACHELOR OF TECHNOLOGY

in

INFORMATION TECHNOLOGY

by

Tanya Gupta
Enrollment No.:04214803114

Archit Gupta
Enrollment No.:04114803114

Guided by

Dr. Bhoomi Gupta
Assistant Professor



MAHARAJA AGRASEN INSTITUTE OF TECHNOLOGY
AFFILIATED TO GURU GOBIND SINGH INDRAPRASTHA UNIVERSITY
SECTOR 22, ROHINI, DELHI-110086

(2014-2018)

STUDYING THE INDUSTRIAL PATTERN USING HADOOP

MAJOR PROJECT REPORT

*Submitted in partial fulfillment of the requirement for the award of the degree
of*

BACHELOR OF TECHNOLOGY

in

INFORMATION TECHNOLOGY

by

Tanya Gupta
Enrollment No.:04214803114

Archit Gupta
Enrollment No.:04114803114

Guided by

Dr. Bhoomi Gupta
Assistant Professor



MAHARAJA AGRASEN INSTITUTE OF TECHNOLOGY
AFFILIATED TO GURU GOBIND SINGH INDRAPRASTHA UNIVERSITY
SECTOR 22, ROHINI, DELHI-110086

(2014-2018)

DECLARATION

It is hereby certified that the work which is being presented in the B. Tech Major Project Report entitled " **Studying the Industrial Pattern using Hadoop** " in partial fulfilment of the requirements for the award of the degree of **Bachelor of Technology** and submitted in the **Department of Information Technology** of **Maharaja Agrasen Institute of Technology, New Delhi (Affiliated to Guru Gobind Singh Indraprastha University, Delhi)** is an authentic record of our own work carried out during a period from **January 2018 to May 2018** under the guidance of **Dr. Bhoomi Gupta, Assistant Professor**.

The matter presented in the B. Tech Major Project Report has not been submitted by me for the award of any other degree of this or any other Institute.

Tanya Gupta
En. No.: 04214803114

Archit Gupta
En. No.: 04114803114

This is to certify that the above statement made by the candidate is correct to the best of my knowledge. They are permitted to appear in the External Major Project Examination

Dr. Bhoomi Gupta
(Assistant Professor)

Dr. M.L Sharma
(HOD, IT)

The B.Tech Minor Project Viva-Voce Examination of **Archit Gupta (04114803114)** and **Tanya Gupta (04214803114)**, has been held on

Project Coordinator

(Signature of External Examiner)

ABSTRACT

This is an era of Big Data. Big Data is driving radical changes in traditional data analysis platforms. To perform any kind of analysis on such voluminous and complex data, scaling up the hardware platforms becomes imminent and choosing the right hardware/software platforms becomes a crucial decision if the user's requirements are to be satisfied in a reasonable amount of time.

Energy Sector of United Kingdom using a very large dataset which represents the list of companies/industries persisting all over the world under different domains. This project is focused on processing such huge dataset using commodity hardware under HADOOP which is otherwise very hard to process using conventional software owing to the size of data. Our project signifies the utility of Hadoop in handling large datasets and furthermore lays out the study of current industrial pattern in UK. Results are warning us about the continuous depletion of nonrenewable resources in the region, specially coal and petroleum resources lying under North Sea. We have laid out some future strategies which can be followed to save these resources and adopt renewable resources to more extent along with further investments in nuclear sector which is very promising source of energy for the region.

Results fetched after analysis are being utilized to plot the structure of Energy sector which would have been costly and difficult using normal processors and memory architecture.

ACKNOWLEDGEMENT

We express our deep gratitude to **Dr. Bhoomi Gupta**, Assistant Professor, Department of Information Technology for his valuable guidance and suggestion throughout my project work. We are thankful to **Nitesh Kumar**, Project Coordinator, for their valuable guidance.

We would like to extend my sincere thanks to **Head of the Department, Dr. M.L Sharma** for his time to time suggestions to complete my project work. We are also very grateful to faculty of the Information Technology Department, Maharaja Agrasen Institute of Technology, Delhi for their ready support.

We wish to express our indebtedness to our parents whose blessings and support always helped us to face the challenges ahead. We are thankful to our friends who were ready to help every time we were stuck with some problem.

Tanya Gupta
En. No.: 04214803114

Archit Gupta
En. No.: 04114803114

CONTENTS

	Page No.
Declaration	iii
Abstract	iv
Acknowledgement	v
Introduction	1
Types of Data	4
Hadoop – Big Data Tool	5
Hadoop Ecosystem	10
Big Data Analysis	12
Code snippets	15
Data Results	22
Analysis	26
Conclusion	30
Future Enhancements	31
References	32

LIST OF FIGURES

S.No.	Fig No.	Screenshot Description	Page Number
01	Figure 1	Big Data Characteristics	01
02	Figure 2	Big Data Customer Scenario	02
03	Figure 3	Hadoop	05
04	Figure 4	Architecture of Hadoop	08
05	Figure 5	HADOOP ECO SYSTEM	10
06	Figure 6	Sqoop	12
07	Figure 7	Graph depicting the results	23
08	Figure 8	Energy sector distribution (UK)	23
09	Figure 9	Distribution global comparison	25
10	Figure 10	Energy sector distribution in 1980	27
11	Figure 11	Energy sector distribution in 2012	27

INTRODUCTION

Big Data

Big data is a popular term used to describe the exponential growth and availability of data, both structured, unstructured and Semi Structured. And big data may be as important to business – and society – as the Internet has become.

- Lots of Data (Terabytes or Gigabytes)
- Big data is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications. The challenges include capture, storage, search, sharing, transfer, analysis, and visualization.
- Systems / Enterprises, Internet users, generate huge amount of data from Gigabytes to and even Terabytes of information.

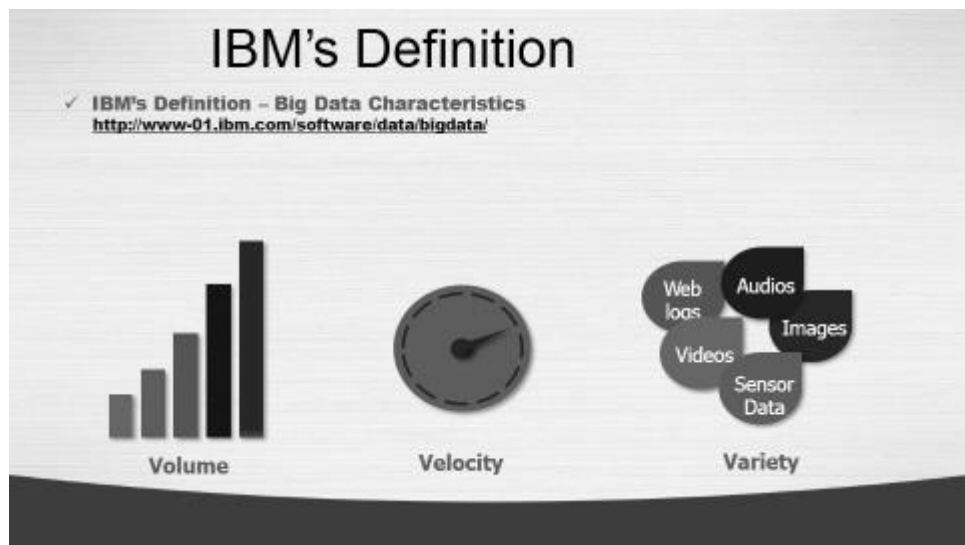


Fig. 1

- **Volume.** Many factors contribute to the increase in data volume. Transaction-based data stored through the years. Unstructured data streaming in from social media. Increasing amounts of sensor and machine-to-machine data being collected. In the past, excessive data volume was a storage issue. But with decreasing storage costs, other issues emerge, including how to determine

relevance within large data volumes and how to use analytics to create value from relevant data.

- **Velocity.** Data is streaming in at unprecedented speed and must be dealt with in a timely manner. RFID tags, sensors and smart metering are driving the need to deal with torrents of data in near-real time. Reacting quickly enough to deal with data velocity is a challenge for most organizations.
- **Variety.** Data today comes in all types of formats. Structured, numeric data in traditional databases. Information created from line-of-business applications. Unstructured text documents, email, video, audio, stock ticker data and financial transactions. Managing, merging and governing different varieties of data is something many organizations still grapple with.

Common Big Data Customer Scenarios

- ✓ **Web and e-tailing**
 - ✓ Recommendation Engines
 - ✓ Ad Targeting
 - ✓ Search Quality
 - ✓ Abuse and Click Fraud Detection
- ✓ **Telecommunications**
 - ✓ Customer Churn Prevention
 - ✓ Network Performance Optimization
 - ✓ Calling Data Record (CDR) Analysis
 - ✓ Analyzing Network to Predict Failure

<http://wiki.apache.org/hadoop/PoweredBy>



Slide

Fig. 2

Big Data Challenges

- **Need for speed** - Now This Time hypercompetitive business environment, companies not only have to find and analyze the relevant data they need, they must find it quickly. Visualization helps organizations perform analyses and make decisions much more rapidly, but the challenge is going through the sheer volumes of data and accessing the level of detail needed, all at a high speed.
- **Data quality** - Analyze data quickly and put it in the proper context for the audience that will be consuming the information, the value of data for decision-making purposes if the data is not accurate or timely. This is a challenge with any data analysis, but when considering the volumes of information involved in big data projects, it becomes even more pronounced.
- **Understanding the data** - It takes a lot of understanding to get data in the right shape so that you can use visualization as part of data analysis. For example, if the data comes from social media content, you need to know who the user is in a general sense – such as a customer using a particular set of products – and understand what it is you're trying to visualize out of the data. Without some sort of context, visualization tools are likely to be of less value to the user.
- **Handling the volume** - The most obvious challenge associated with big data is simply storing and analyzing all that information. In its Digital Universe report, IDC estimates that the amount of information stored in the world's IT systems is doubling about every two years. By 2020, the total amount will be enough to fill a stack of tablets that reaches from the earth to the moon 6.6 times. And enterprises have responsibility or liability for about 85 percent of that information.

TYPES OF DATA

Unstructured Data

Unstructured data files often include text and multimedia content. Examples include e-mail messages, word processing documents, videos, photos, audio files, presentations, webpages and many other kinds of business documents. While these sorts of files may have an internal structure, they are still considered "unstructured" because the data they contain doesn't fit neatly in a database. Experts estimate that 80 to 90 percent of the data in any organization is unstructured. And the amount of unstructured data in enterprises is growing significantly often many times faster than structured databases are growing.

Unstructured data is all those things that can't be so readily classified and fit into a neat box like photos and graphic images, videos, streaming instrument data, webpages, PDF files, PowerPoint presentations, emails, blog entries, wikis and word processing documents.

Structured data

Data that resides in a fixed field within a record or file is called structured data. This includes data contained in relational databases and spreadsheets.

Structured data first depends on creating a data model – a model of the types of business data that will be recorded and how they will be stored, processed and accessed. This includes defining what fields of data will be stored and how that data will be stored: data type (numeric, currency, alphabetic, name, date, address).

Semi-Structured Data

Semi-structured data is a cross between the two. It is a type of structured data, but lacks the strict data model structure. For eg, XML.

HADOOP –BIGDATA TOOL

Introduction to Hadoop

Apache Hadoop is an open-source software framework written in Java for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware. All the modules in Hadoop are designed with a fundamental assumption that hardware failures (of individual machines, or racks of machines) are commonplace and thus should be automatically handled in software by the framework.

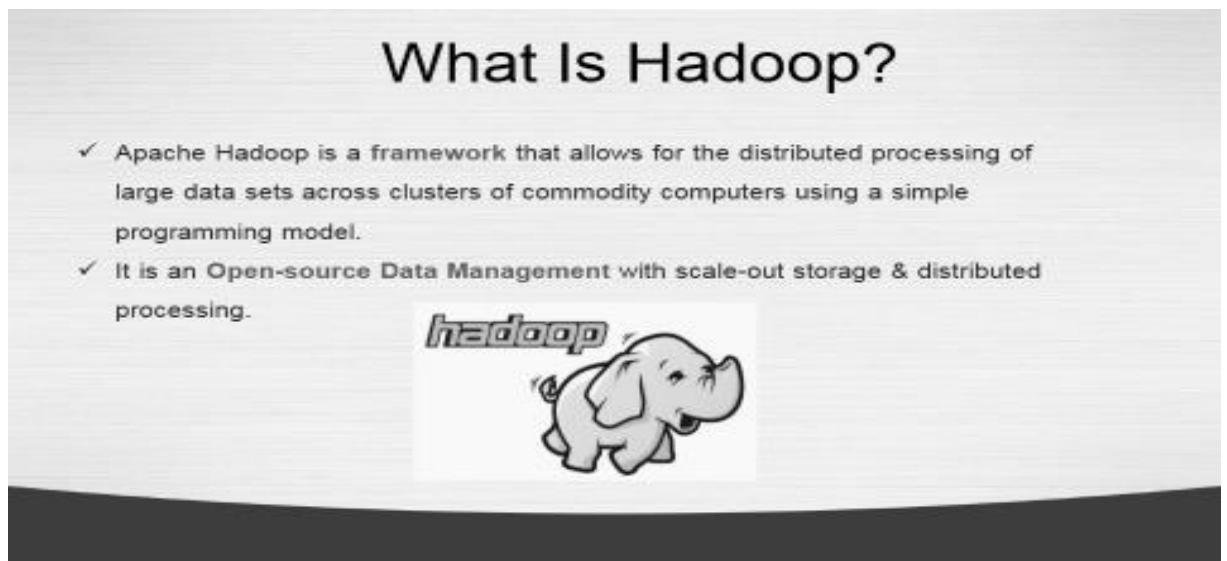


Fig. 3

- Hadoop provides a reliable shared storage (HDFS) and analysis system (MapReduce).
- Hadoop is highly scalable and unlike the relational databases, Hadoop scales linearly. Due to linear scale, a Hadoop Cluster can contain tens, hundreds, or even thousands of servers.
- Hadoop is very cost effective as it can work with commodity hardware and does not require expensive high-end hardware.

History of Hadoop

Google invented the basic frameworks that constitute what is today popularly called as Hadoop. They faced the future first with problem of handling billions of searches and indexing millions of web pages. When they couldn't find any large scale, distributed, scalable computing platforms for their needs, they just went ahead and created their own.

Doug Cutting was inspired by Google's white papers and decided to create an open source project called "Hadoop".

Doug Cutting, Cloudera's Chief Architect, helped create Hadoop out of necessity as data from web exploded, and grew far beyond the ability of traditional systems to handle it.

Yahoo further contributed to this project and played a key role in developing Hadoop for enterprise applications. Since then many companies such as Facebook, LinkedIn, ebay, Horton works, Cloudera etc have contributed to the Hadoop project.

Advantages of Hadoop

- ✓ **Data Size and Data Diversity** - When you are dealing with huge volumes of data coming from various sources and in a variety of formats then you can say that you are dealing with Big Data. In this case, Hadoop is the right technology for you.
- ✓ **Future Planning** - It is all about getting ready for challenges you may face in future. If you anticipate Hadoop as a future need then you should plan accordingly. To implement Hadoop on you data you should first understand the level of complexity of data and the rate with which it is going to grow. So, you need a cluster planning. It may begin with building a small or medium cluster in your industry as per data (in GBs or few TBs) available at present and scale up your cluster in future depending on the growth of your data.

- ✓ **Multiple Frameworks for Big Data** - There are various tools for various purposes. Hadoop can be integrated with multiple analytic tools to get the best out of it, like Mahout for Machine-Learning, R and Python for Analytics and visualization, Python, Spark for real time processing, MongoDB and Hbase for No sql database, Pentaho for BI etc.
- ✓ **Lifetime Data Availability** - When you want your data to be live and running forever, it can be achieved using Hadoop's scalability. There is no limit to the size of cluster that you can have. You can increase the size anytime as per your need by adding data nodes to it with minimal cost.
- ✓ **Cost Effective** - Hadoop also offers a cost effective storage solution for businesses' exploding data sets. The problem with traditional relational database management systems is that it is extremely cost prohibitive to scale to such a degree in order to process such massive volumes of data. In an effort to reduce costs, many companies in the past would have had to down-sample data and classify it based on certain assumptions as to which data was the most valuable. The raw data would be deleted, as it would be too cost-prohibitive to keep. While this approach may have worked in the short term, this meant that when business priorities changed, the complete raw data set was not available, as it was too expensive to store.
- ✓ **Scalable** - Hadoop is a highly scalable storage platform, because it can store and distribute very large data sets across hundreds of inexpensive servers that operate in parallel. Unlike traditional relational database systems (RDBMS) that can't scale to process large amounts of data, Hadoop enables businesses to run applications on thousands of nodes involving many thousands of terabytes of data.
- ✓ **Some Other Uses**
 - Log file processing
 - Analysis of Text, Image, Audio, & Video content
 - Recommendation systems like in E-Commerce Websites

Architecture of Hadoop

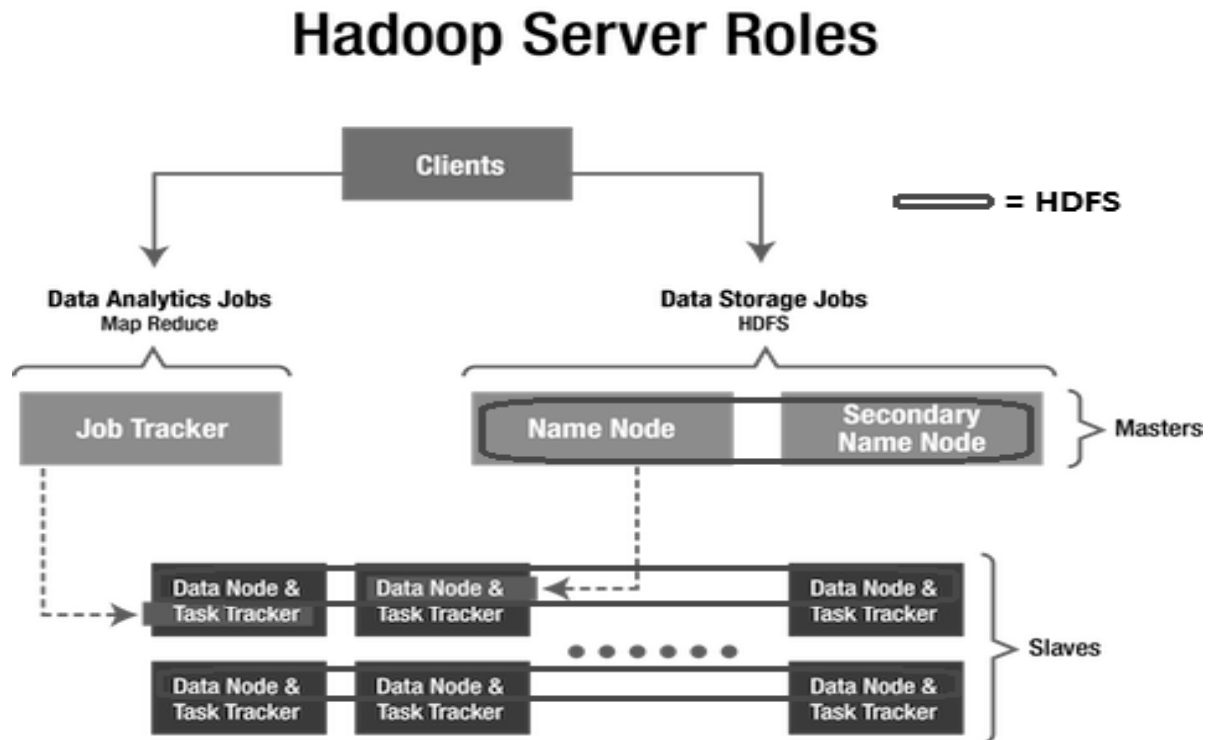


Fig. 4

- Hadoop works in a master-worker / master-slave fashion.
- Hadoop has two core components: HDFS and MapReduce.
- **HDFS (Hadoop Distributed File System)** offers a highly reliable and distributed storage, and ensures reliability, even on a commodity hardware, by replicating the data across multiple nodes. Unlike a regular file system, when data is pushed to HDFS, it will automatically split into multiple blocks (configurable parameter) and stores/replicates the data across various datanodes. This ensures high availability and fault tolerance. HDFS can be deployed on a broad spectrum of machines that support Java. Though one can run several DataNodes on a single machine, but in the practical world, these DataNodes are spread across various machines.

- **MapReduce** offers an analysis system which can perform complex computations on large datasets. This component is responsible for performing all the computations and works by breaking down a large complex computation into multiple tasks and assigns those to individual worker/slave nodes and takes care of coordination and consolidation of results. For processing large sets of data MR comes into the picture. The programmers will write MR applications that could be suitable for their business scenarios. Programmers have to understand the MR working flow and according to the flow, applications will be developed and deployed across Hadoop clusters. Hadoop built on Java APIs and it provides some MR APIs that is going to deal with parallel computing across nodes.
- The master contains the Namenode and Job Tracker components.
 - **Namenode** holds the information about all the other nodes in the Hadoop Cluster, files present in the cluster, constituent blocks of files and their locations in the cluster, and other information useful for the operation of the Hadoop Cluster.
 - **Job Tracker** keeps track of the individual tasks/jobs assigned to each of the nodes and coordinates the exchange of information and results.
- Each Worker / Slave contains the Task Tracker and a Datanode components.
 - **Task Tracker** is responsible for running the task / computation assigned to it.
 - **Datanode** is responsible for holding the data.
- The computers present in the cluster can be present in any location and there is no dependency on the location of the physical server.

HADOOP ECO SYSTEM

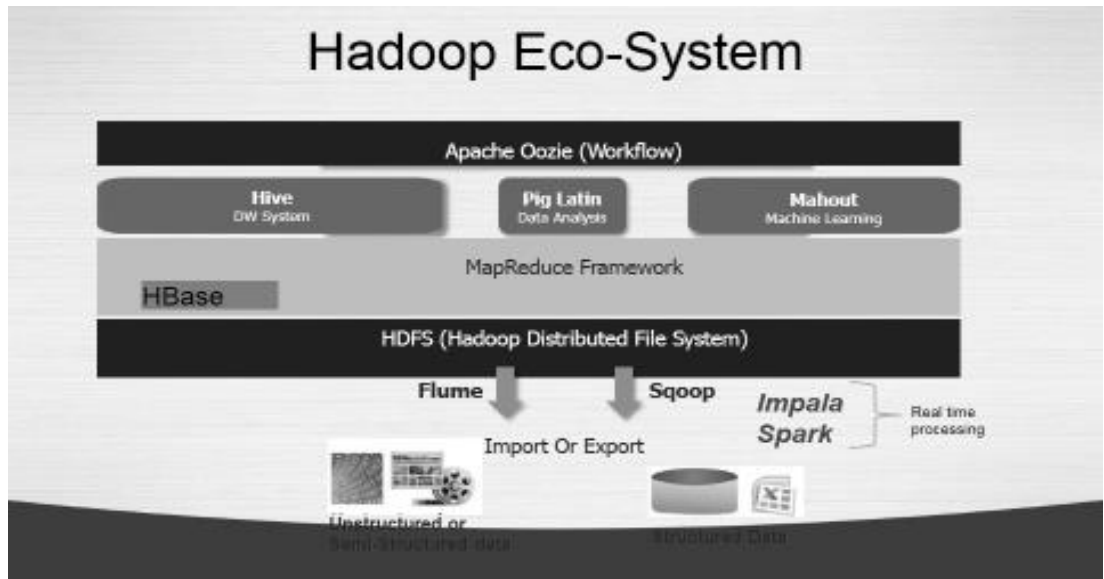


Fig 5

HIVE

Hive provides a warehouse structure and SQL-like access for data in HDFS and other Hadoop input sources (e.g. Amazon S3). Hive's query language, HiveQL, compiles to MapReduce. It also allows user defined functions (UDFs). Hive is widely used, and has itself become a "sub-platform" in the Hadoop ecosystem.

PIG

Pig is a framework consisting of a high-level scripting language (Pig Latin) and a run-time environment that allows users to execute MapReduce on a Hadoop cluster. Like HiveQL in Hive, Pig Latin is a higher-level language that compiles to MapReduce.

MAHOUT

Mahout is a scalable machine-learning and data mining library. There are currently four main groups of algorithms in Mahout:

- recommendations, also known as collective filtering

- classification, also known as categorization
- clustering
- frequent item set mining, also known as parallel frequent pattern mining

MapReduce

- The MapReduce paradigm for parallel processing comprises two sequential steps: map and reduce.
- In the map phase, the input is a set of key-value pairs and the desired function is executed over each key/value pair in order to generate a set of intermediate key/value pairs.

HBASE

Based on Google's Bigtable, HBase "is an open-source, distributed, versioned, column-oriented store" that sits on top of HDFS. HBase is column-based rather than row-based, which enables high-speed execution of operations performed over similar values across massive data sets, e.g. read/write operations that involve all rows but only a small subset of all columns. HBase does not provide its own query or scripting language, but is accessible through Java, Thrift, and REST APIs.

SQOOP

Sqoop ("SQL-to-Hadoop") is a tool which transfers data in both directions between relational systems and HDFS or other Hadoop data stores, e.g. Hive or HBase.

According to the Sqoop blog, "You can use Sqoop to import data from external structured datastores into Hadoop Distributed File System or related systems like Hive and HBase. Conversely, Sqoop can be used to extract data from Hadoop and export it to external structured datastores such as relational databases and enterprise data warehouses."

BIG DATA ANALYSIS

Using Sqoop and MySql and HDFS we can handle the data and Transfer the data MySql to HDFS and analyse the Data.

Working of Sqoop

Sqoop is a tool designed to transfer data between Hadoop and relational database servers. It is used to import data from relational databases such as MySQL, Oracle to Hadoop HDFS, and export from Hadoop file system to relational databases. It is provided by the Apache Software Foundation. SQL to Hadoop and Hadoop to SQL. Through Sqoop we can Import full table, part of table and selected value into Hadoop.

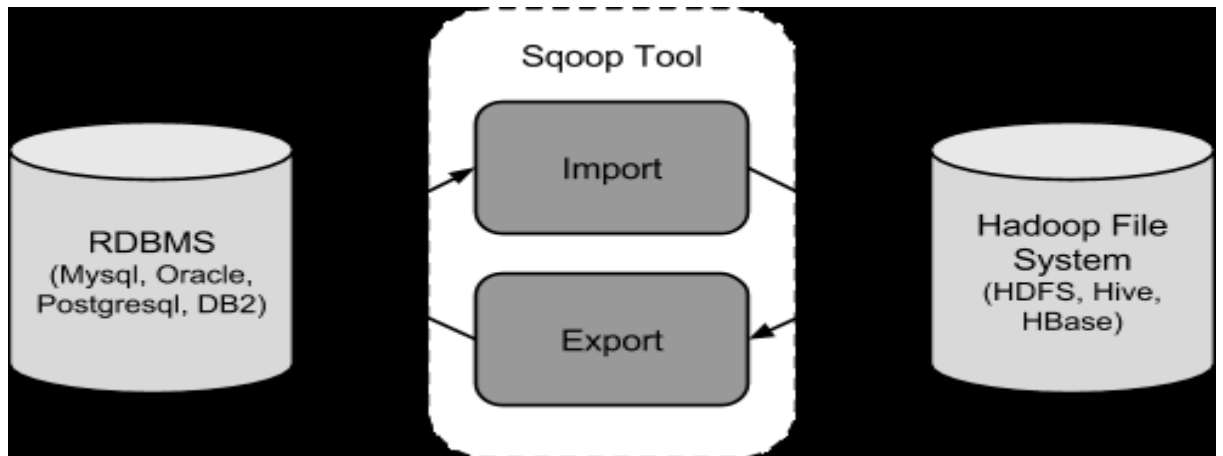


Fig. 6

- **Sqoop Import** - The import tool imports individual tables from RDBMS to HDFS. Each row in a table is treated as a record in HDFS. All records are stored as text data in text files or as binary data in Avro and Sequence files.
- **Sqoop Export** - The export tool exports a set of files from HDFS back to an RDBMS. The files given as input to Sqoop contain records, which are called as rows in table. Those are read and parsed into a set of records and delimited with user-specified delimiter

Steps To Install and Create Table on Operating system

- ✓ On Centos or Ubuntu, install MySQL.
- ✓ Create Database.
- ✓ Create Table schema into MYSQL.
- ✓ Insert the values into table one by one or Dump full data into MYSQL Table.
- ✓ Now your full data into your table.

Let us take an example of 1tables named as **H_info** which are in a database called **HadoopGyan** in a MySQL database server.

H_info

ID	Name	Country
101	Map reduce	USA
102	Pig	UK
103	Hive	India
104	HBase	Africa

Importing Full Table to HDFS

This string will connect to a MySQL database named Hadoopgyan. The connect string will be used on TaskTracker nodes throughout MapReduce cluster; if specify the literal name localhost, each node will connect to a different database (or more likely, no database at all). When full hostname or IP address of the database host were used they were seen by all remote nodes.

If not authenticate against the database before you can access it, then can use the --username and --password or -P parameters to supply a username and a password to the database.

Sqoop automatically supports several databases, including MySQL. Connect strings beginning with jdbc:mysql:// are handled automatically in Sqoop.

Sqoop Import

- `$ Sqoop Import --connect jdbc:mysql://local host/Database name -- username -
- table name --target-dir /HDFS -M 1`
- `$ Sqoop Import --connect jdbc:mysql://local host/ Hadoopgyan --username root
--table H_info --Target-dir /user/cloudera/HadoopGyan -- m 1`

Through above command our table H_info import to the hadoop (HDFS)

Sqoop Export

Command to Export data Into MYSQL From HDFS this is called Export in SQOOP

- `Sqoop export --connect jdbc:mysql://localhost:portnumber/databasename --
username root -P --table name --export-dir "/path of table "`
- Or
- `$ sqoop export --connect jdbc:mysql://localhost/Portnumber/databasename --
username root -Password Password --table name --export-dir "Path of table"`

CODE SNIPPLETS & SCREENSHOTS

putting data on hdfs architecture

```
hadoop fs -put /home/cloudera/Desktop/Data.csv /user/cloudera/newdata.txt;
```

showing Data file on host

```
http://localhost:50070/explorer.html#/user/cloudera
```

opening pig

```
pig -x local;
```

Loading full data in a variable

```
variable = Load '/home/cloudera/Desktop/Data.csv' using PigStorage(',') as  
(CompanyName:chararray,CompanyNumber:int,RegAddressCareOf:chararray,RegAdd  
ressPOBox:chararray,RegAddressAddressLine1:chararray,RegAddressAddressLine2:ch  
ararray,RegAddressPostTown:chararray,RegAddressCounty:chararray,RegAddressCou  
ntry:chararray,RegAddressPostCode:chararray,CompanyCategory:chararray,CompanyS  
tatus:chararray,CountryOfOrigin:chararray,DissolutionDate:chararray,IncorporationDat  
e:chararray,AccountsAccountRefDay:int,AccountsAccountRefMonth:chararray,Accoun  
tsNextDueDate:chararray,AccountsLastMadeUpDate:chararray,AccountsAccountCateg  
ory:chararray>ReturnsNextDueDate:chararray>ReturnsLastMadeUpDate:chararray,Mort  
gagesNumMortCharges:int,MortgagesNumMortOutstanding:int,MortgagesNumMortPa  
rtSatisfied:chararray,MortgagesNumMortSatisfied:chararray,SICCodeSicText_1:chararr  
ay,SICCodeSicText_2:chararray,SICCodeSicText_3:chararray,SICCodeSicText_4:char  
array,LimitedPartnershipsNumGenPartners:int,LimitedPartnershipsNumLimPartners:int  
,URI:chararray,PreviousName_1CONDATE:chararray,PreviousName_1CompanyName  
:chararray,PreviousName_2CONDATE:chararray,PreviousName_2CompanyName:char  
array,PreviousName_3CONDATE:chararray,PreviousName_3CompanyName:chararra  
y,PreviousName_4CONDATE:chararray,PreviousName_4CompanyName:chararray,Pr  
eviousName_5CONDATE:chararray,PreviousName_5CompanyName:chararray,Previo  
usName_6CONDATE:chararray,PreviousName_6CompanyName:chararray,PreviousN  
ame_7CONDATE:chararray,PreviousName_7CompanyName:chararray,PreviousName
```

```
_8CONDATE:chararray,PreviousName_8CompanyName:chararray,PreviousName_9CONDATE:chararray,PreviousName_9CompanyName:chararray,PreviousName_10CONDATE:chararray,PreviousName_10CompanyName:chararray,ConfStmtNextDueDate:chararray,ConfStmtLastMadeUpDate:chararray);
```

loading csv file into a variable

```
var = Load '/home/cloudera/Desktop/data.csv' using PigStorage(',') as  
(Date:chararray,name:chararray,SIC1:chararray,SIC2:chararray,SIC3:chararray,SIC4:chararray,  
address1:chararray,address2:chararray,address3:chararray,address4:chararray,address5:chararray);
```

displaying the variable

```
dump var;
```

#filter global data for different SIC codes

```
fillglobal = filter var by (SIC1 == '05101 - Deep coal mines' or SIC2 == '05101 - Deep coal mines' or SIC3 == '05101 - Deep coal mines' or SIC4 == '05101 - Deep coal mines');
```

#counting the filtered out dates

```
countglobal = foreach (group fillglobal all) generate COUNT(fillglobal);
```

#display count

```
Dump countglobal;
```

#filter uk data

```
filluk = filter var by (SIC1 == '05101 - Deep coal mines' or SIC2 == '05101 - Deep coal mines' or SIC3 == '05101 - Deep coal mines' or SIC4 == '05101 - Deep coal mines') and  
(address1 == 'UNITED KINGDOM' or address2 == 'UNITED KINGDOM' or address3 == 'UNITED KINGDOM' or address4 == 'UNITED KINGDOM' or address5 == 'UNITED KINGDOM');
```

#counting the filtered out dates

countuk = foreach (group filluk all) generate COUNT(filluk);

#display count

Dump countuk;

converting the dates into datetime format

convertuk = foreach filluk generate ToDate(Date, 'dd-mm-yyyy') as (ukDate:datetime);

displaying the dates

dump convertuk;

filtering the dates in a particular range > year 2005

fillukdate = filter convertuk1 by ukDate >= (datetime)ToDate('2005-01-01', 'yyyy-mm-dd');

displaying the filtered out dates

dump filldate;

counting the filtered out dates

counteruk = foreach (group fillukdate all) generate COUNT(fillukdate);

displaying the count

dump counteruk;

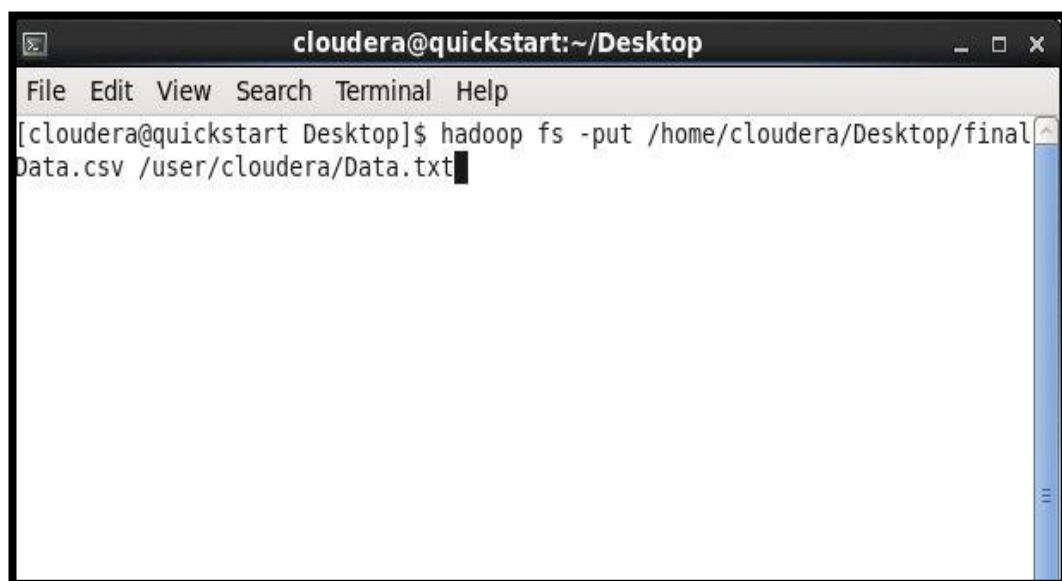
Open Cloudera CDH for Analyzing Big Data

Cloudera is a commodity software for Hadoop which works on linux for which a software emulator – VMWare is used.



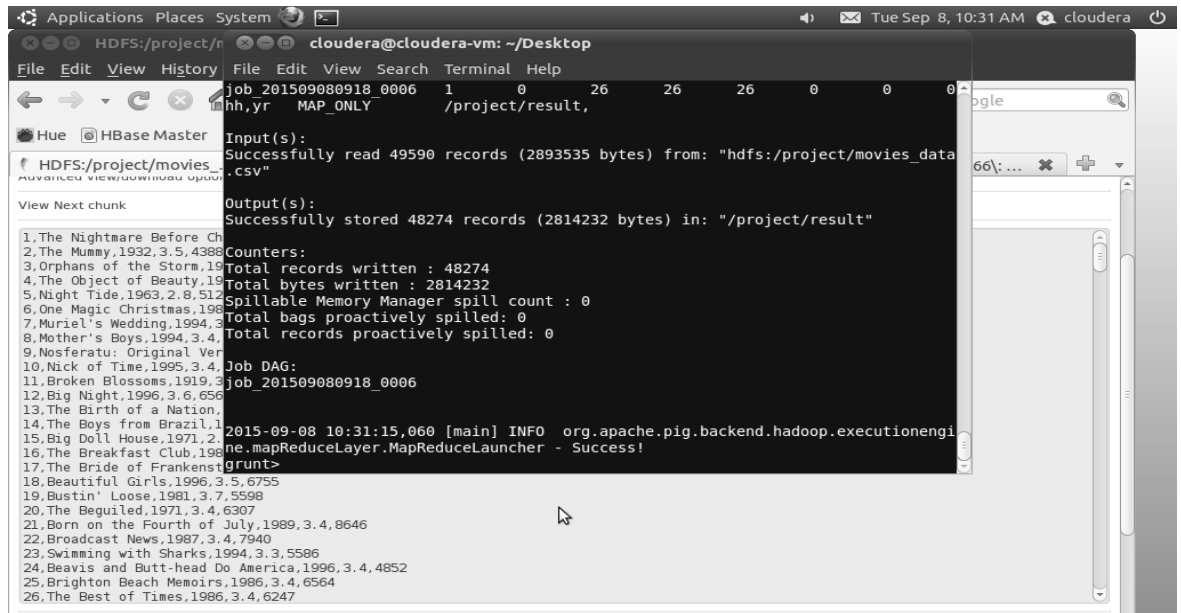
Copy Input File into HDFS

The data file is copied onto HDFS architecture using linux command –put.



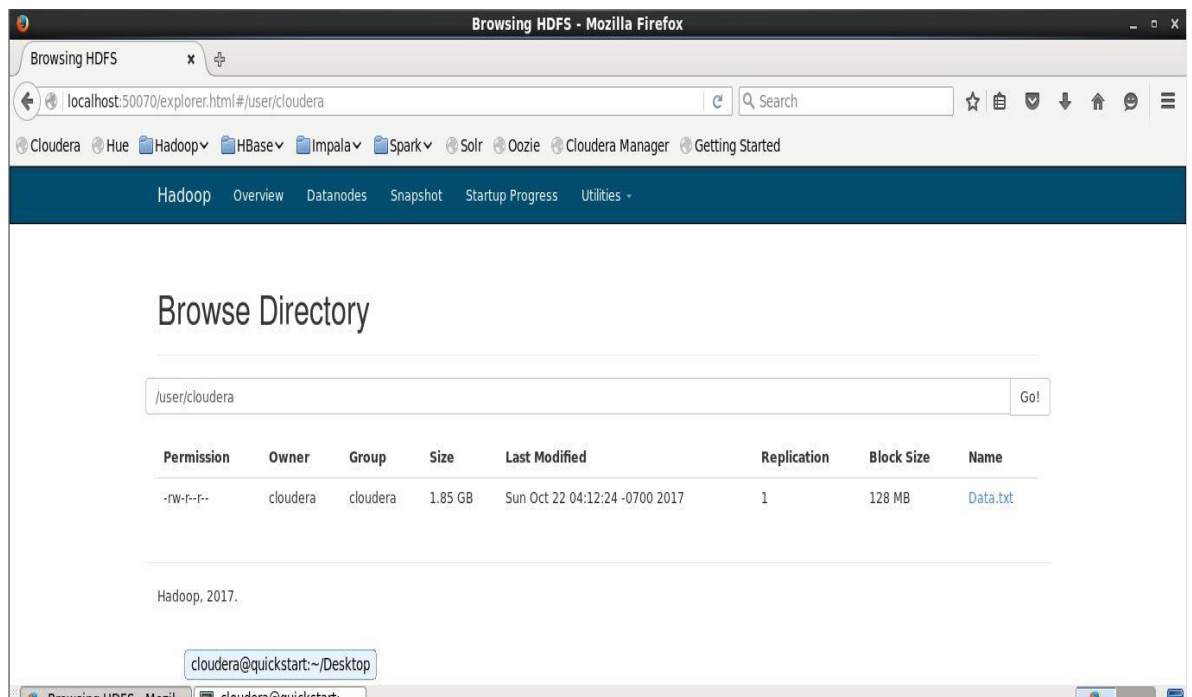
Copy Data into HDFS

The data file is being copied onto HDFS architecture memory whose progress can be seen in the following screenshot.



Resulted Data into HDFS

The copied data file can finally be viewed on HDFS architecture memory browsing through the relevant directory. Respective permissions can be seen on the extreme left. Size of the big data file can be seen under the size tab. Here we have used a file of 2 GB size.



Create Schema in Pig

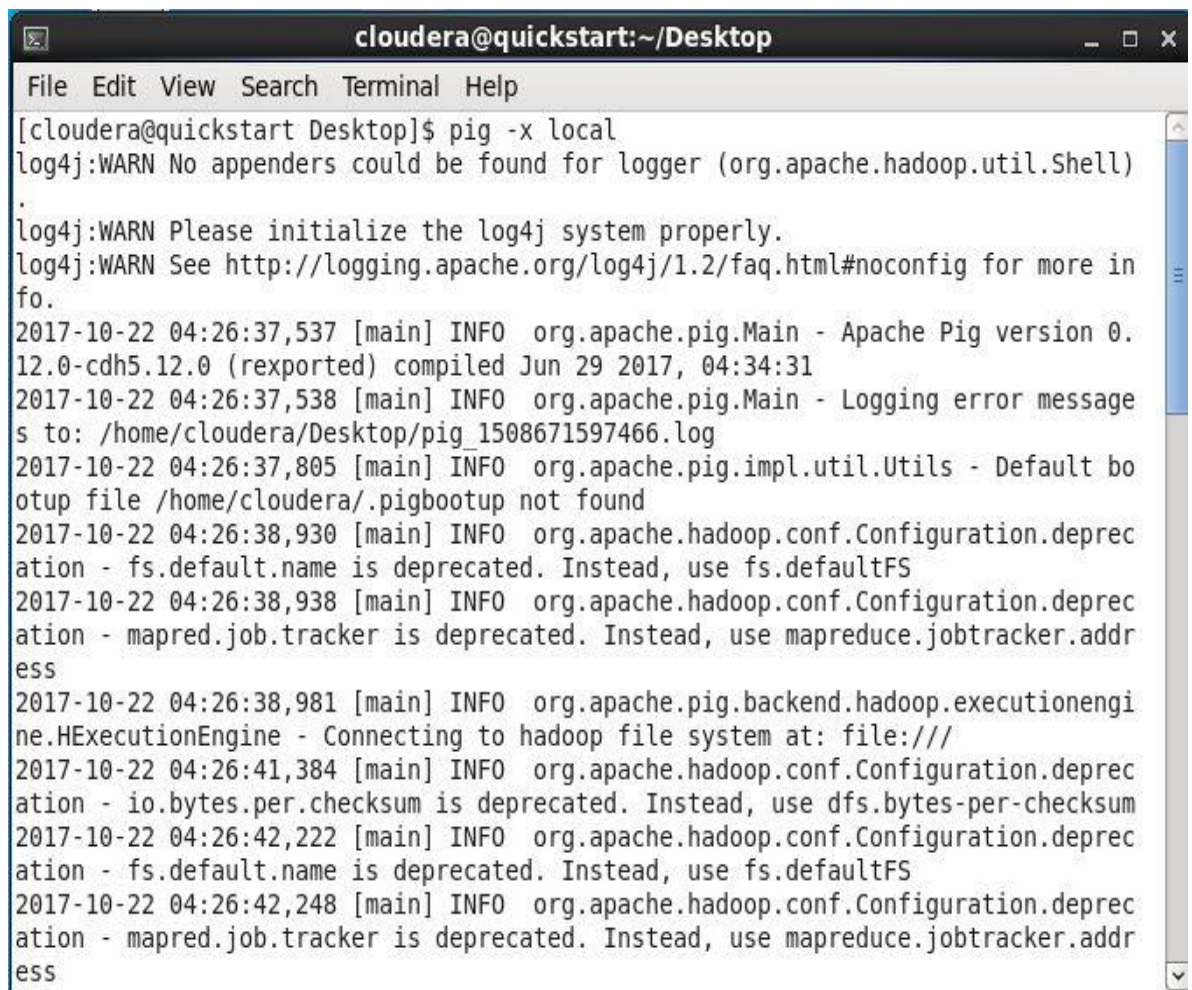
Linux command prompt is used to use a scripting language named PIG

Pig is a framework consisting of a high-level scripting language (Pig Latin) and a run-time environment that allows users to execute MapReduce on a Hadoop cluster.

Like HiveQL in Hive, Pig Latin is a higher-level language that compiles to MapReduce.

To enable pig scripting, following command is used.

Pig scripting is done in Linux Command Prompt window.



```
cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help
[cloudera@quickstart Desktop]$ pig -x local
log4j:WARN No appenders could be found for logger (org.apache.hadoop.util.Shell)
.
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more in
fo.
2017-10-22 04:26:37,537 [main] INFO  org.apache.pig.Main - Apache Pig version 0.
12.0-cdh5.12.0 (rexpoted) compiled Jun 29 2017, 04:34:31
2017-10-22 04:26:37,538 [main] INFO  org.apache.pig.Main - Logging error message
s to: /home/cloudera/Desktop/pig_1508671597466.log
2017-10-22 04:26:37,805 [main] INFO  org.apache.pig.impl.util.Utils - Default bo
otup file /home/cloudera/.pigbootup not found
2017-10-22 04:26:38,930 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2017-10-22 04:26:38,938 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2017-10-22 04:26:38,981 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.HExecutionEngine - Connecting to hadoop file system at: file:///
2017-10-22 04:26:41,384 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2017-10-22 04:26:42,222 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2017-10-22 04:26:42,248 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
```

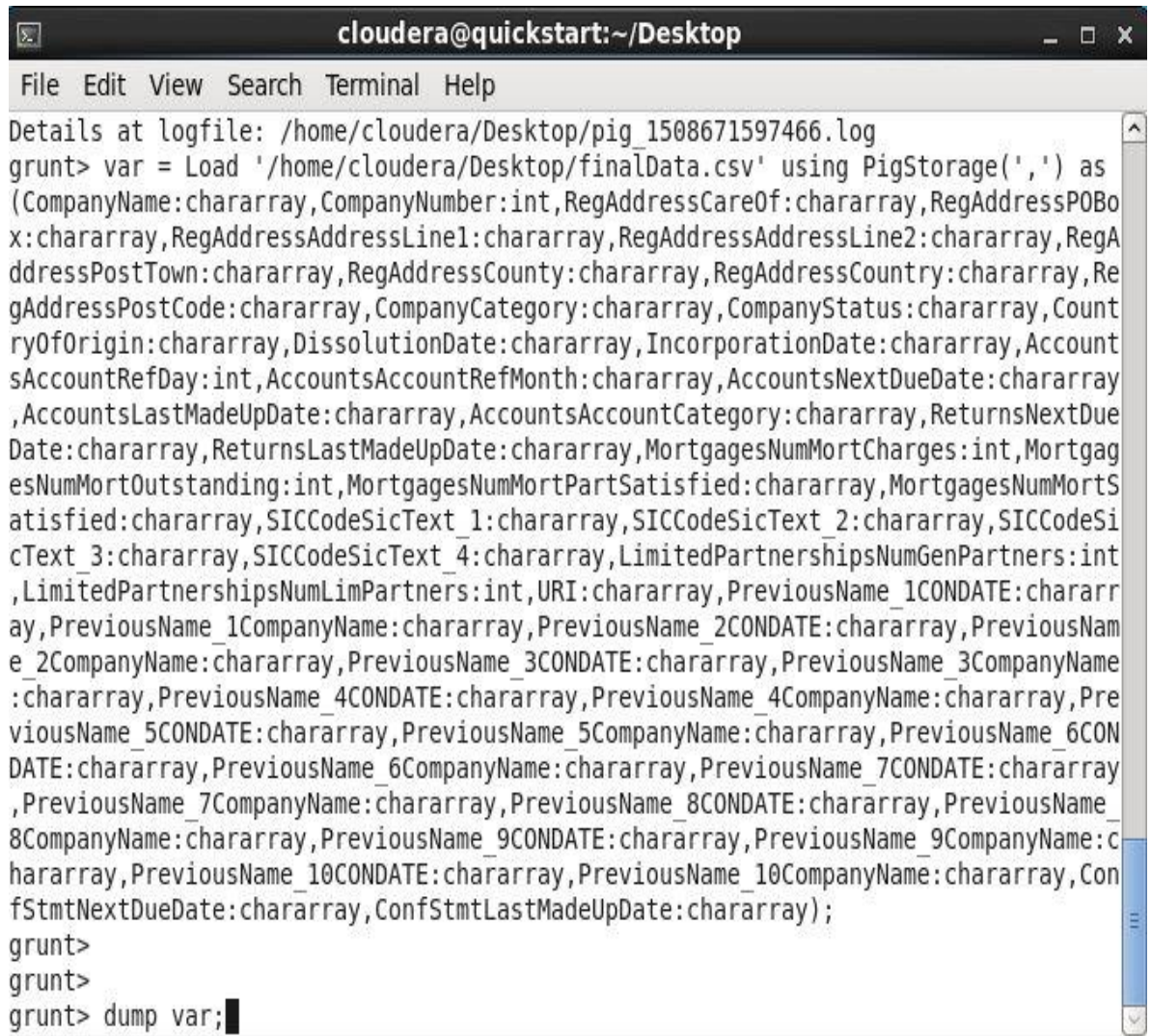
Loading data in a variable in Pig Schema

To work with the data file, it is first loaded into a variable.

Command “Load” is used to load the data in the temporary storage.

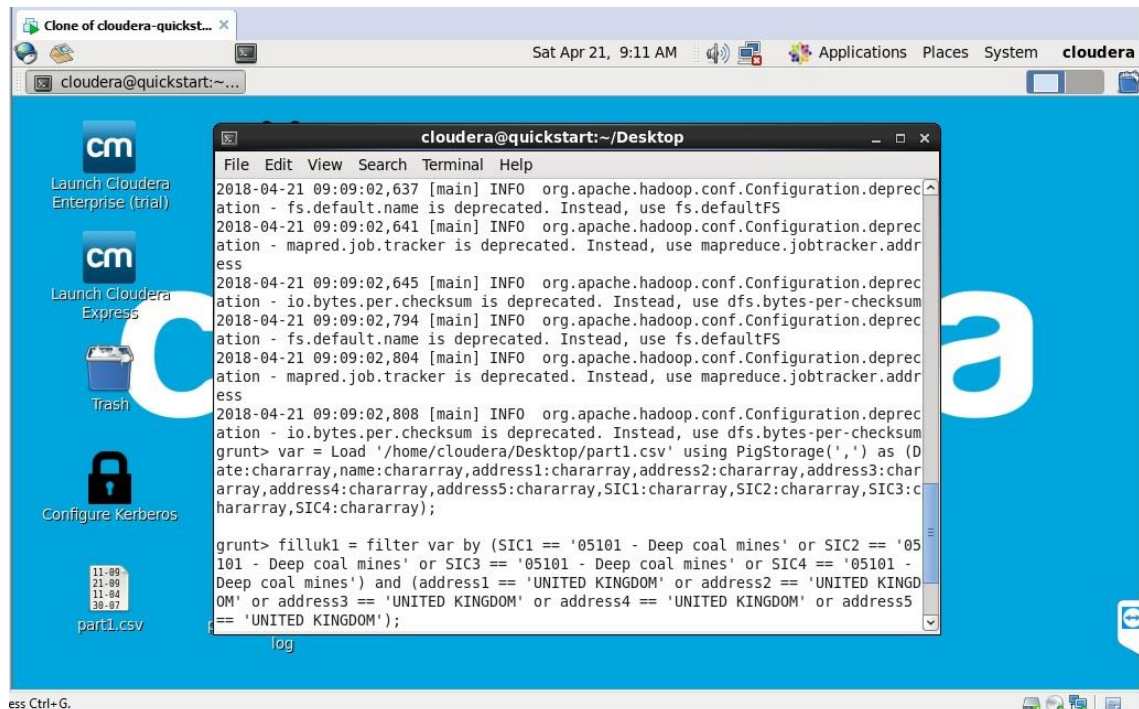
After Load, the path of the file is entered.

Each column in the csv file is denoted with a column name in the Load command.



```
cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help
Details at logfile: /home/cloudera/Desktop/pig_1508671597466.log
grunt> var = Load '/home/cloudera/Desktop/finalData.csv' using PigStorage(',') as
(CompanyName:chararray,CompanyNumber:int,RegAddressCareOf:chararray,RegAddressPOBo
x:chararray,RegAddressAddressLine1:chararray,RegAddressAddressLine2:chararray,RegA
ddressPostTown:chararray,RegAddressCounty:chararray,RegAddressCountry:chararray,Re
gAddressPostCode:chararray,CompanyCategory:chararray,CompanyStatus:chararray,Count
ryOfOrigin:chararray,DissolutionDate:chararray,IncorporationDate:chararray,Account
sAccountRefDay:int,AccountsAccountRefMonth:chararray,AccountsNextDueDate:chararray
,AccountsLastMadeUpDate:chararray,AccountsAccountCategory:chararray>ReturnsNextDue
Date:chararray>ReturnsLastMadeUpDate:chararray,MortgagesNumMortCharges:int,Mortgag
esNumMortOutstanding:int,MortgagesNumMortPartSatisfied:chararray,MortgagesNumMorts
atisfied:chararray,SICCodeSicText_1:chararray,SICCodeSicText_2:chararray,SICCodeSi
cText_3:chararray,SICCodeSicText_4:chararray,LimitedPartnershipsNumGenPartners:int
,LimitedPartnershipsNumLimPartners:int,URI:chararray,PreviousName_1CONDATE:chararr
ay,PreviousName_1CompanyName:chararray,PreviousName_2CONDATE:chararray,PreviousNam
e_2CompanyName:chararray,PreviousName_3CONDATE:chararray,PreviousName_3CompanyName
:chararray,PreviousName_4CONDATE:chararray,PreviousName_4CompanyName:chararray,Pre
viousName_5CONDATE:chararray,PreviousName_5CompanyName:chararray,PreviousName_6CON
DATE:chararray,PreviousName_6CompanyName:chararray,PreviousName_7CONDATE:chararray
,PreviousName_7CompanyName:chararray,PreviousName_8CONDATE:chararray,PreviousName_
8CompanyName:chararray,PreviousName_9CONDATE:chararray,PreviousName_9CompanyName:c
hararray,PreviousName_10CONDATE:chararray,PreviousName_10CompanyName:chararray,Con
fStmtNextDueDate:chararray,ConfStmtLastMadeUpDate:chararray);
grunt>
grunt>
grunt> dump var;
```


Fetching data for SIC code 05101 - Deep coal mines in UK



The screenshot shows a Cloudera desktop environment with a terminal window titled "cloudera@quickstart:~/Desktop". The terminal displays the following output:

```
File Edit View Search Terminal Help
2018-04-21 09:09:02,637 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-04-21 09:09:02,641 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2018-04-21 09:09:02,645 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2018-04-21 09:09:02,794 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-04-21 09:09:02,804 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2018-04-21 09:09:02,808 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
grunt> var = Load '/home/cloudera/Desktop/part1.csv' using PigStorage(',') as (D
ate:chararray,name:chararray,address1:chararray,address2:chararray,address3:char
array,address4:chararray,address5:chararray,SIC1:chararray,SIC2:chararray,SIC3:c
hararray,SIC4:chararray);

grunt> filluk1 = filter var by (SIC1 == '05101 - Deep coal mines' or SIC2 == '05
101 - Deep coal mines' or SIC3 == '05101 - Deep coal mines' or SIC4 == '05101 -
Deep coal mines') and (address1 == 'UNITED KINGDOM' or address2 == 'UNITED KING
DOM' or address3 == 'UNITED KINGDOM' or address4 == 'UNITED KINGDOM' or address5
== 'UNITED KINGDOM');

log
```

Result for above query



The screenshot shows a Cloudera desktop environment with a terminal window titled "cloudera@quickstart:~/Desktop". The terminal displays the following output:

```
File Edit View Search Terminal Help
Output(s):
Successfully stored records in: "file:/tmp/temp1280263871/tmp822551836"

Job DAG:
job_local1318967827_0001

2018-04-21 09:12:46,043 [main] INFO org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Success!
2018-04-21 09:12:46,067 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-04-21 09:12:46,068 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2018-04-21 09:12:46,070 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2018-04-21 09:12:46,071 [main] WARN org.apache.pig.data.SchemaTupleBackend - Sc
hemaTupleBackend has already been initialized
2018-04-21 09:12:46,437 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileI
nputFormat - Total input paths to process : 1
2018-04-21 09:12:46,438 [main] INFO org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(1)
grunt>
```

Data fetched successfully

```
cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help
2018-04-21 09:15:31,648 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2018-04-21 09:15:31,649 [main] INFO org.apache.pig.tools.pigstats.SimplePigStats - Detected Local mode. Stats reported below may be incomplete
2018-04-21 09:15:31,650 [main] INFO org.apache.pig.tools.pigstats.SimplePigStats - Script Statistics:

HadoopVersion  PigVersion      UserId      StartedAt      FinishedAt      Features
2.6.0-cdh5.12.0 0.12.0-cdh5.12.0 cloudera    2018-04-21 09:15:31 2018-04-21 09:15:18 2
018-04-21 09:15:31 GROUP_BY,FILTER

Success!

Job Stats (time in seconds):
JobId  Alias  Feature  Outputs
job_local1514965119_0002 1-35,convertuk2,counteruk2,filluk2,fillukdate2,var
ar      GROUP_BY,COMBINER      file:/tmp/temp1280263871/tmp1662051737,

Input(s):
Successfully read records from: "/home/cloudera/Desktop/part1.csv"

Output(s):
Successfully stored records in: "file:/tmp/temp1280263871/tmp1662051737"
```

Results for similar queries are shown below

```
cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help
Output(s):
Successfully stored records in: "file:/tmp/temp1280263871/tmp-2087125955"

Job DAG:
job_local2060627611_0003

2018-04-21 09:17:15,362 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-04-21 09:17:15,363 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-04-21 09:17:15,363 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2018-04-21 09:17:15,364 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2018-04-21 09:17:15,364 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-04-21 09:17:15,391 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-04-21 09:17:15,391 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(83)
grunt>
```



```
cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help
Output(s):
Successfully stored records in: "file:/tmp/temp1280263871/tmp-1938685507"

Job DAG:
job_local1801859764_0004

2018-04-21 09:19:40,644 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-04-21 09:19:40,646 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-04-21 09:19:40,648 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2018-04-21 09:19:40,652 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2018-04-21 09:19:40,659 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-04-21 09:19:40,716 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-04-21 09:19:40,716 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(37)
grunt> █
```

```
cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help
Successfully stored records in: "file:/tmp/temp1280263871/tmp702324092"

Job DAG:
job_local415324901_0005

2018-04-21 09:21:04,915 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-04-21 09:21:04,919 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-04-21 09:21:04,919 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2018-04-21 09:21:04,921 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2018-04-21 09:21:04,923 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-04-21 09:21:04,969 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-04-21 09:21:04,969 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
grunt>
grunt> █
```

```
cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help
Successfully stored records in: "file:/tmp/temp1280263871/tmp-585196303"

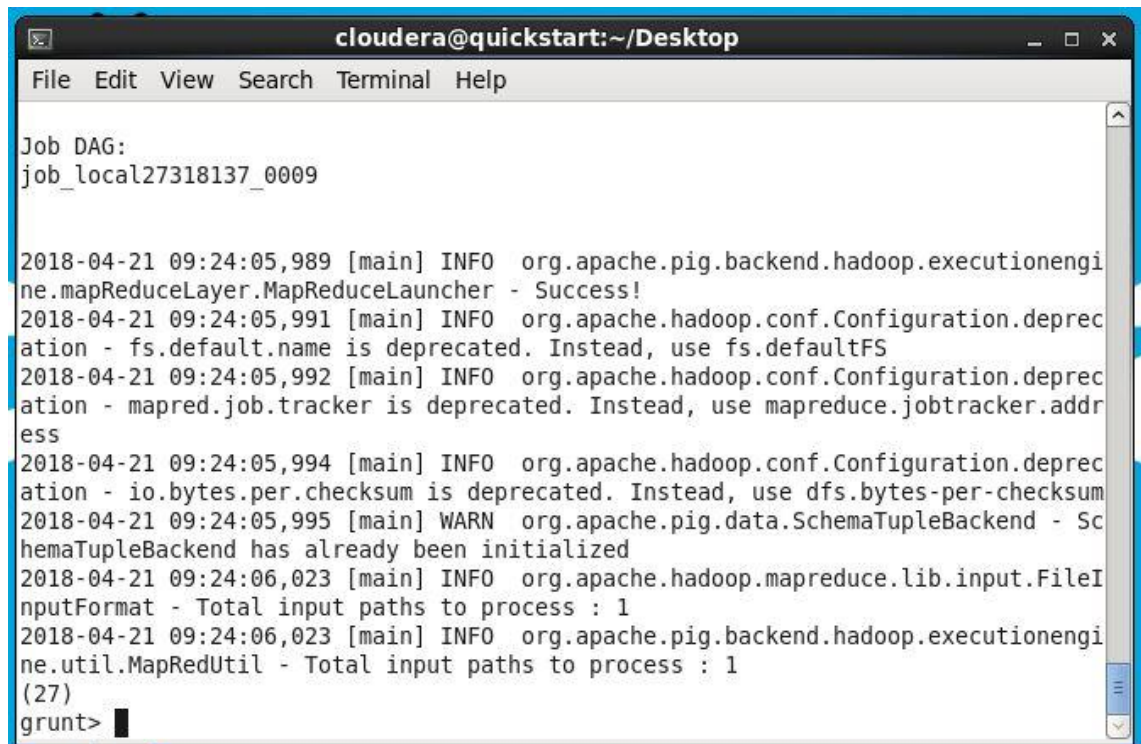
Job DAG:
job_local192217008_0006

2018-04-21 09:21:55,353 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-04-21 09:21:55,355 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-04-21 09:21:55,355 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2018-04-21 09:21:55,357 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2018-04-21 09:21:55,358 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-04-21 09:21:55,399 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-04-21 09:21:55,399 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(154)
grunt>
```

```
cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help
Successfully stored records in: "file:/tmp/temp1280263871/tmp-176476672"

Job DAG:
job_local2046776129_0007

2018-04-21 09:22:47,261 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2018-04-21 09:22:47,262 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-04-21 09:22:47,262 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2018-04-21 09:22:47,263 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2018-04-21 09:22:47,263 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-04-21 09:22:47,279 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-04-21 09:22:47,279 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(3)
grunt>
```

```
cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help

Job DAG:
job_local27318137_0009

2018-04-21 09:24:05,989 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Success!
2018-04-21 09:24:05,991 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-04-21 09:24:05,992 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2018-04-21 09:24:05,994 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2018-04-21 09:24:05,995 [main] WARN  org.apache.pig.data.SchemaTupleBackend - Sc
hemaTupleBackend has already been initialized
2018-04-21 09:24:06,023 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileI
nputFormat - Total input paths to process : 1
2018-04-21 09:24:06,023 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(27)
grunt> █
```

DATA RESULTS

The Dataset of companies which is huge in size is first accessed through Cloudera software using VMware virtual machine.

To get results from the data, we use PIG script to fire various queries on the data.

Linux command shell is used in Cloudera to implement the queries.

In our project, we focused on the energy sector on a global and country level with the United Kingdom as our first test case. Under the energy sector, we found the relevant SIC codes related to find the industries associated with that SIC code.

We tested the following SIC codes:

05101 - Deep coal mines

06100 - Extraction of crude petroleum

06200 - Extraction of natural gas

07210 - Mining of uranium and thorium ores

09100 - Support activities for petroleum and natural gas extraction

20110 - Manufacture of industrial gases

24460 - Processing of nuclear fuel

46711 - Wholesale of petroleum and petroleum products

After testing all the SIC codes, we reached to following results:

Energy Sectors	Number of Industries (Global)	Number of Industries (United Kingdom)
Coal	145	16
Crude	2061	295
Natural Gas	967	162
Uranium & Thorium	14	3
Petroleum & Natural Gas	3935	719
Industrial Gases	52	6
Nuclear Fuel	11	2
Wholesale of Petroleum	510	107

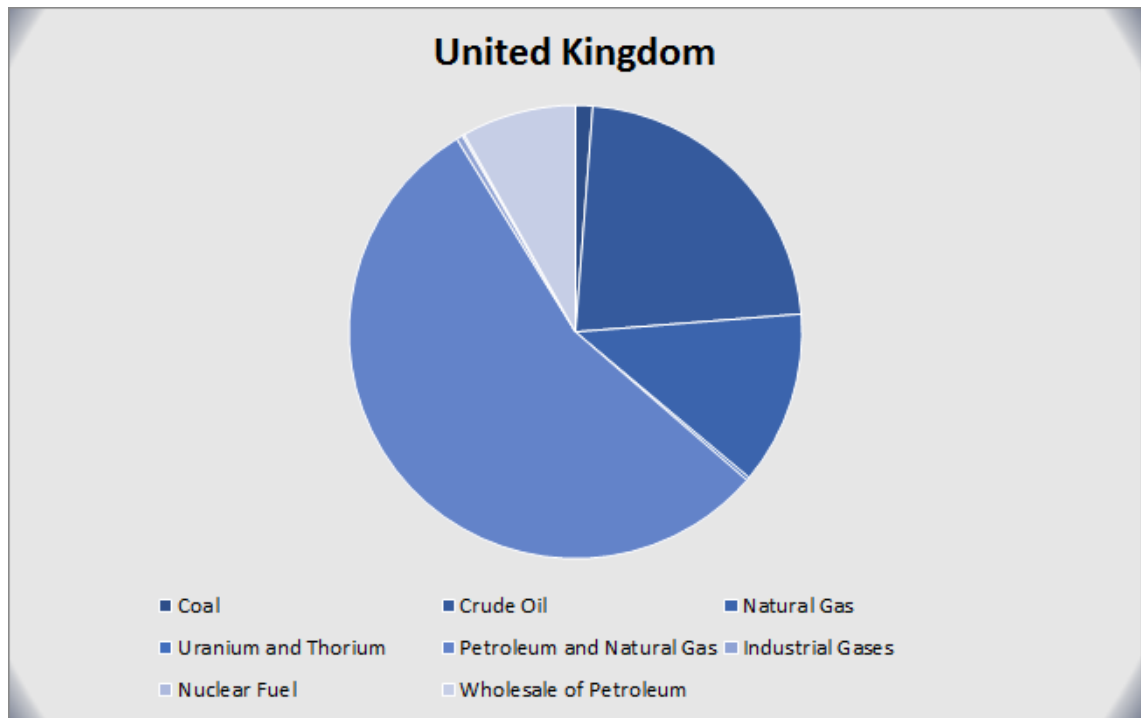


Fig. 7

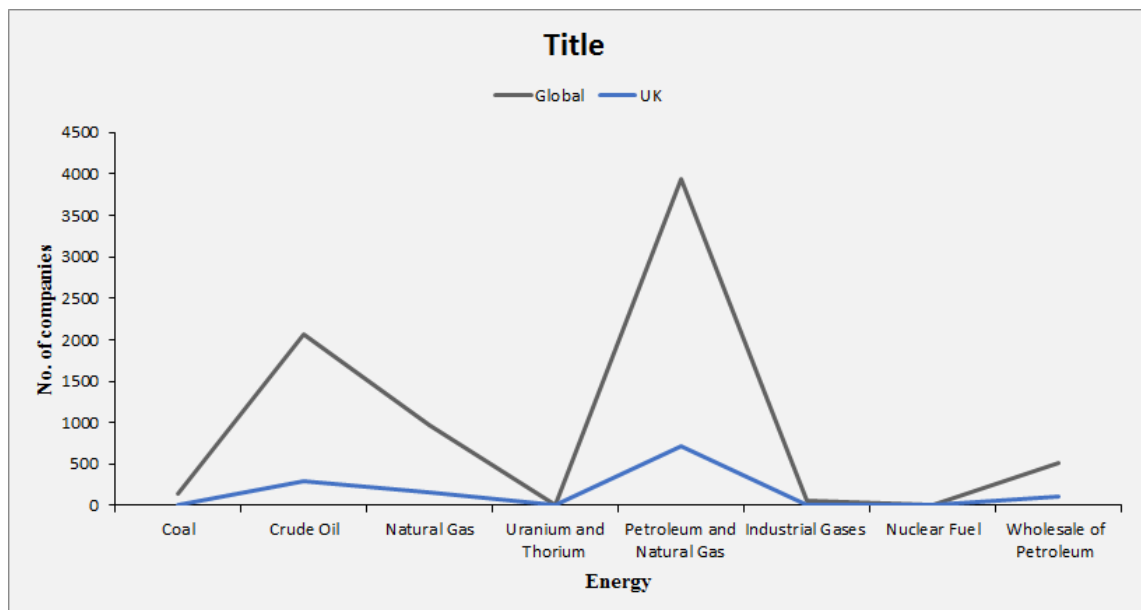


Fig. 8

The United Kingdom has always found to be endowed with energy resources. Our results clearly show a major portion of world's industry sector lying in United Kingdom. While the graph contains the combined results of number of industries being set up from early 1900s till date, the United Kingdom never lacked to be a significant producer of energy, especially in petroleum sector as we can see from the graph, among

3935 industries, almost 700 are located in the United Kingdom which is quite big number.

The United Kingdom used to be the largest producer of coal historically but its coal industry started declining significantly after World War I which is the reason for its low contribution in coal industry at present. On the other hand, our results still show a significant participation of the UK alone in coal industries among the world owing to the 16 number of industries belonging in the UK form 145 all over the world.

Energy distribution inside the UK

Our pie chart shows the energy distribution of the United Kingdom at present. Petroleum and natural gas are the prime area of work for the UK in today's time. This major shift of industrial sector from coal during early 20th century to petroleum and natural gas during the mid-20th century was due to discovery of oil supplies in the North Sea which lies under the United Kingdom. The petroleum sector provided energy to almost 10% of the UK which is quite big percentage compared to other countries.

Strategies to promote Renewable Energy

We can presume several possibilities for the United Kingdom from our results. As we know, petroleum is non-renewable source of energy and can lead to significant decline in resources if continued to exploit. Therefore the United Kingdom needs to gradually shift to renewable sources of energy which gives need to adopt following strategies –

- Replace coal with natural gas, which can import in the form of liquefied natural gas –Through this strategy, the country will succeed and its environmental objectives.
- Increasing the role of nuclear power and renewable energy sources in the future energy mix which will significantly reduce the need to exploit coal & petroleum.

Variation since 20th Century

After comparing the industrial sector of the United Kingdom at global level, we further dig into the data to see the growth of energy sector in the United Kingdom since the 20th century and we came to following conclusion after covering major sectors such as

COAL

PETROLEUM AND NATURAL GAS

NUCLEAR FUEL

Energy Sectors	Number of Industries (United Kingdom)				
	1900-2000	2000-2005	2005-2010	2010-2015	2015-2018
COAL	3	0	0	4	9
CRUDE	19	13	18	34	256
NATURAL GAS	8	3	10	18	144
NUCLEAR FUEL	0	0	2	0	4
PETROLEUM	25	21	31	79	759

Compiling these results in graphical form, we get –

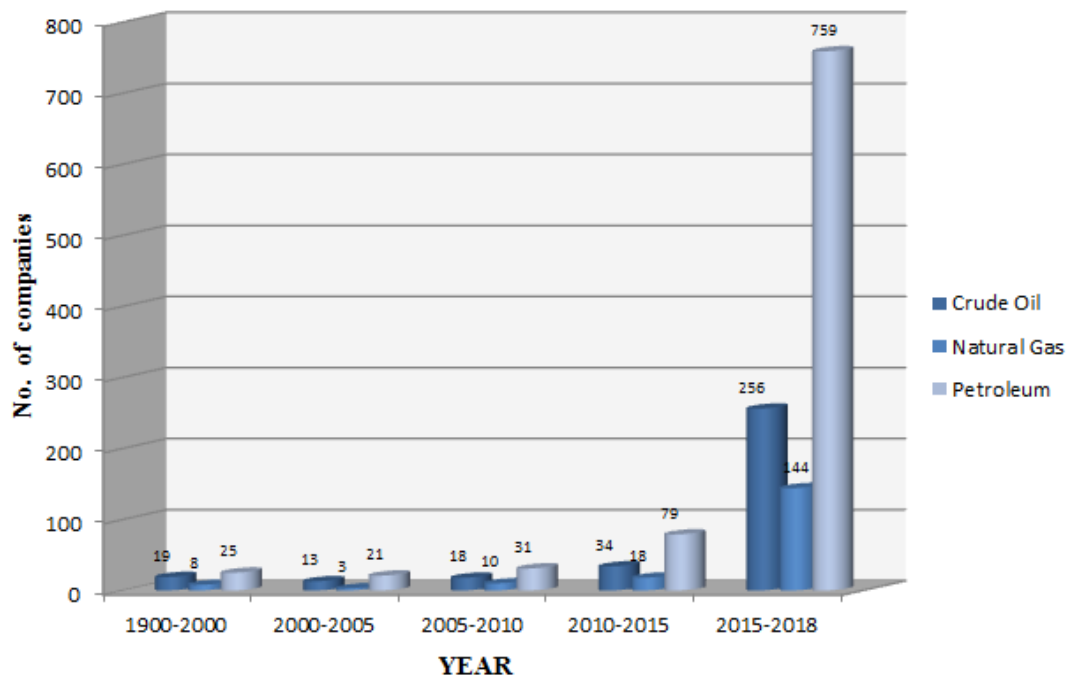


Fig. 9

ANALYSIS

The results are quite interesting as they are showing several ups and downs for various energy sectors in the United Kingdom.

As we saw in previous analysis, the United Kingdom is richly endowed with energy resources. Historically, the country relied on coal mining and only tentatively ventured into nuclear energy in the mid-1950s. In the 1960s, the U.K. turned to the oil and natural gas buried below the North Sea. Although the country's natural resources are decreasing, the production of primary energy still accounts for 10% of Britain's Gross Domestic Product (GDP), a much higher share than in the majority of industrialized countries.

In its history, the United Kingdom was majorly dependent on coal for its production, but in the late 20th century, the profit from coal industry declined significantly as the history says that is the reason behind no growth in coal industry since the 21st century.

Moreover, analysis says that further exploitation of coal could have resulted in significant depletion of the resource.

Our results show a slight interest in nuclear energy sector at start of the 21st century. This change was brought up to shift to renewable sources of energy in the United Kingdom as constant exploitation of petroleum for energy needs was constantly depleting the resources.

Coming on the petroleum sector, this sector contains several ups and downs. History says late 1900s was the peak time of production of petroleum in United Kingdom. But the growth became steady at the start of 21st century due to heavy production during 1999-2000. As the graph shows, the petroleum sector, although steadily, kept growing in the United Kingdom to supply energy needs in the country and continued to play the highest percentage in the country among energy sources.

The following figures can be used to compare the growth of different sectors from late 1900s to 21st century which matches our results.

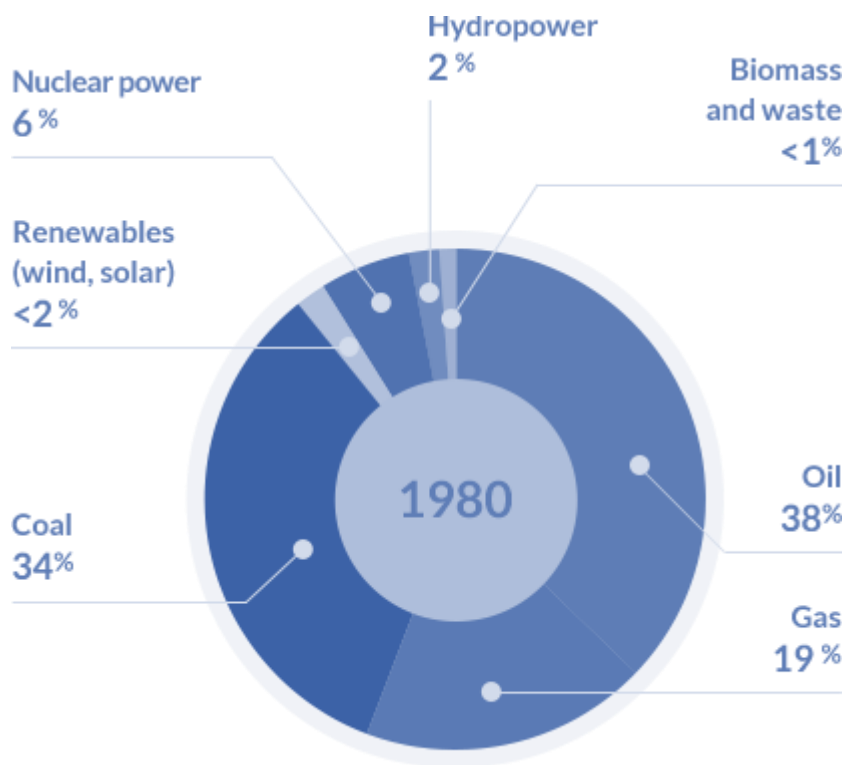


Fig. 10

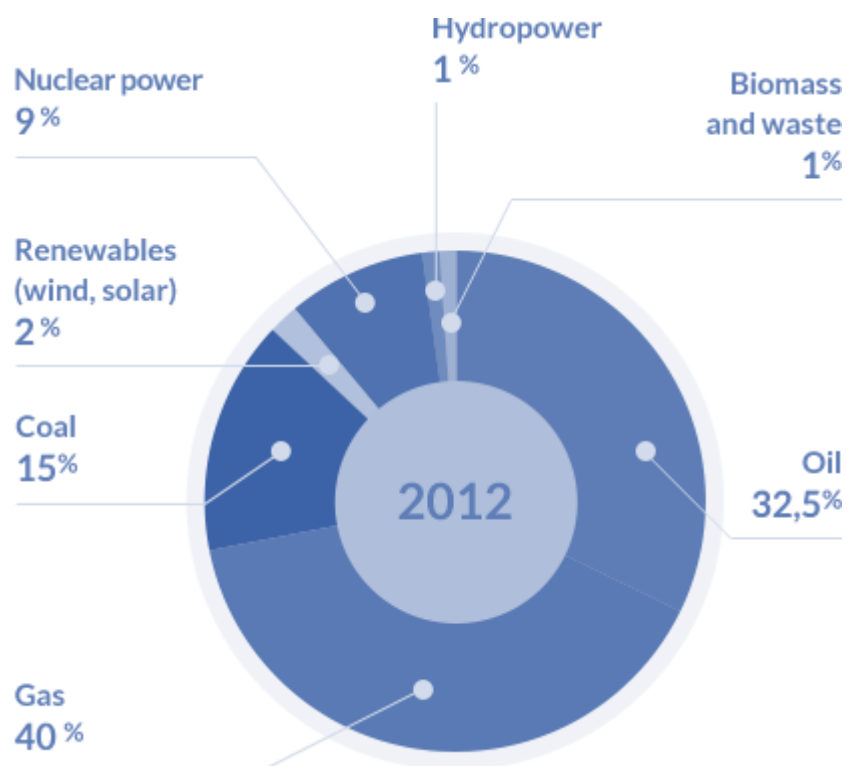
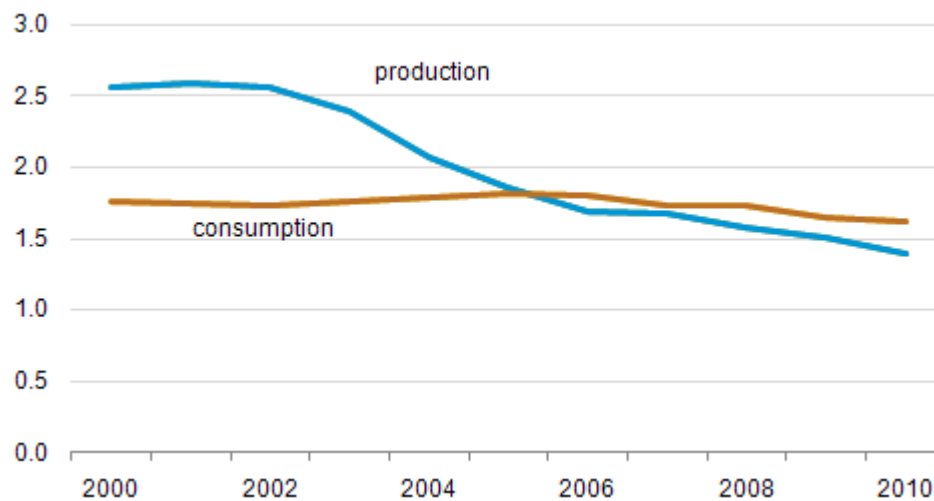


Fig. 11

As we can see, the use of coal as energy source decreased significantly while natural gas and oil became the prime source of energy in the recent years. This comparison verifies the result shown in the previous graph which showed a high increase in number of industries in petroleum sector while coal industry failed to flourish in that time.

What does our Analysis suggest for future scope in the United Kingdom?



The United Kingdom is the largest producer of oil and second-largest producer of natural gas in the European Union which can also be seen from our results. Due to steadily declining production since the early 2000s, the U.K. became a net importer of natural gas and oil. This decline can be verified by slow growth of industries in petroleum sector in recent years which would have been drastic if the oil supplies would have been surplus.

Despite decreasing production, the U.K. remains one of the European Union's leading petroleum exporters which we signified in our previous analysis.

Because discoveries of new oil and natural gas reserves have not outpaced the maturation of existing oil and natural gas fields, production from both has declined. The U.K.'s increasing reliance on imported natural gas and oil has spurred the government to develop energy policies to focus on enhanced oil and gas recovery, as well as increased cooperation with Norway—U.K.'s largest oil supplier. The U.K. has also invested

heavily in renewable energy; according to the U.K. Department of Energy and Climate Change, the U.K. has the largest offshore wind resource in the world.

Furthermore our analysis suggests following strategies for the United Kingdom to follow to sustain energy requirements in the country.

Continued delays in delivering nuclear and a lack of investors willing to come forward to fund new large gas generating plant could provide opportunities for a range of renewable energy technologies to be deployed at scale if policy is supportive. This will particularly be the case for technologies such as solar and wind, whose costs continue to fall, such that as the end of decade approaches renewable energy plant gets built through investor choice rather than Government policy. The arrival of competitive energy storage solutions and the continued uptake of EVs, which also work at their best when deployed with renewable energy, will be a catalyst for accelerated deployment.

The energy industry is itself divided over the transition to a low carbon economy. Diversification of the energy market, and the providers that supply it, is clearly a pressure point for the main players. If we move to a renewable world, then this will have a mix of large- and small-scale projects. This means that more companies will be able to enter the market creating more competition from the likes of renewable energy companies such as Ecotricity and Good Energy.

Big energy companies have yet to step up to the challenges of a low-carbon future. Despite some impressive investments in renewables, especially in wind and biomass, renewables are still very much periphery activities for most of the major utilities. Renewables should be at the heart of any plans for low-carbon energy generation because they are the only proven low-carbon technologies that are actually getting cheaper.

A lot of future investment can come from new market entrants, such as farmers, businesses, and communities investing in on-site and local renewable energy generation, but we need to take the utilities with us as well.

Combining the scale of utilities with the renewable energy projects, a lot of investment can be made in this sector, especially in the United Kingdom which will affect the whole world in a positive way.

CONCLUSION

Big Data is an evolving field, where much of the research is yet to be done. Big data at present, and is handled by the software named Hadoop. In this project, we utilized the computation power of Hadoop over Big Data to fetch faster results from a data of over 2 GB to analyze some results from the data processing which over normal processors could have taken a lot of time than we took in this project.

The United Kingdom is the largest producer of oil and second-largest producer of natural gas in the European Union which can also be seen from our results. Due to steadily declining production since the early 2000s, the U.K. became a net importer of natural gas and oil. This decline can be verified by the slow growth of industries in the petroleum sector in recent years which would have been drastic if the oil supplies would have been surplus.

Despite decreasing production, the U.K. remains one of the European Union's leading petroleum exporters which we signified in our previous analysis. The U.K. has also invested in renewable energy. Continued delays in delivering nuclear and a lack of investors willing to come forward to fund new large gas generating plant could provide opportunities for a range of renewable energy technologies to be deployed at scale if the policy is supportive. If we move to a renewable world, then this will have a mix of large- and small-scale projects. This means that more companies will be able to enter the market creating more competition from the likes of renewable energy companies. A lot of future investment can come from new market entrants, such as farmers, businesses, and communities investing in on-site and local renewable energy generation.

FUTURE ENHANCEMENT

The world is becoming data driven. Each and every decision is now taken on data from stock markets to machines that stalk you. The power of Hadoop can further be utilized to fetch deeper results from the current data used in the project and even much heavier data which is almost impossible to do normally.

The United Kingdom needs to gradually shift to renewable sources of energy which gives need to adopt following strategies—

- Replace coal with natural gas, which can import in the form of liquefied natural gas- Through this strategy, the country will succeed and its environmental objectives.
- Increasing the role of nuclear power and renewable energy sources in the future energy mix which will significantly reduce the need to exploit coal & petroleum.

REFERENCES

- [1] Apache Hive. Available at <http://hive.apache.org>
- [2] Oracle and FSN, "Mastering Big Data: CFO Strategies to Transform Insight into Opportunity".
- [3] "IBM What is big data? – Bringing big data to the enterprise". www.ibm.com.
- [4] "Big Data for Good" (PDF). ODBMS.org. 5 June 2012.
- [5] Data Set - <https://data.gov.uk/dataset/basic-company-data>
- [6] Final Output of Data - http://download.companieshouse.gov.uk/en_output.html
- [7] Detailed view of dataset - <http://webarchive.nationalarchives.gov.uk/20140711134125/http://www.companieshouse.gov.uk/toolsToHelp/pdf/freeDataProductDataset.pdf>
- [8] <https://www.planete-energies.com/en/medias/saga-energies/history-energy-united-kingdom>
- [9] <https://www.eia.gov/todayinenergy/detail.php?id=3170>
- [10] <https://www.r-e-a.net/blog/overview-of-the-uks-renewable-energy-growth-08-06-2016>
- [11] <https://www.lynda.com/Hadoop-tutorials/Data-Analysis-Hadoop/460439-2.html>
- [12] <https://www.3pillarglobal.com/insights/analyze-big-data-hadoop-technologies>
- [13] <http://www.informit.com/articles/article.aspx?p=2008905>
- [14] <https://hortonworks.com/hadoop-tutorial/how-to-analyze-machine-and-sensor-data/>
- [15] https://www.sas.com/en_in/insights/big-data/hadoop.html
- [16] <https://dzone.com/articles/data-analysis-using-apache-hive-and-apache-pig>
- [17] <https://www.udemy.com/learn-how-to-analyse-hadoop-data-using-apache-pig/>
- [18] <https://www.dezyre.com/article/difference-between-pig-and-hive...hadoop.../79>
- [19] <https://www.ironsidegroup.com/2015/12/01/hadoop-ecosystkey-components/>
- [20] <https://blog.eduonix.com/bigdata-and-hadoop/learn-process-data-using-apache-pig-hadoop-platform/>
- [21] <https://www.energy-uk.org.uk/energy-industry.html>
- [22] <https://www.energy-uk.org.uk/energy-industry/the-energy-market.html>
- [23] https://en.wikipedia.org/wiki/Electricity_sector_in_the_United_Kingdom
- [24] <https://www.theguardian.com/business/energy-industry>

- [25] <https://www.brighetwork.co.uk/career-path-guides/energy.../growth-jobs-energy/>
- [26] D. P. Acharjya, Kauser Ahmed P, “A Survey on Big Data Analytics: Challenges, Open Research Issues and Tools”, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 2, 2016.
- [27] Ramesh Sharda, Daniel Adomako Asamoah and Natraj Ponna, “Research and Pedagogy in Business Analytics: Opportunities and Illustrative Examples”, Journal of Computing and Information Technology - CIT 21, 2013, 3, 171–183 doi:10.2498/cit.1002194.
- [28] Anurag K. Srivastava, Sukumar Kamalasadan, Daxa Patel, Sandhya Sankar Khalid S. Al-Olimat” Electricity markets: an overview and comparative study”, International Journal Of Energy Sector Management, volume 5, issue 2.
- [29] Dr. D.K. Subramanian, Dr. T.V. Ramachandra,” Energy Utilisation in Karnataka -Part II: Industries Sector (IISc.)”, Humanity Development Library Version - 2.0.
- [30] Athanasios Kolios and George Read, “A Political, Economic, Social, Technology, Legal and Environmental (PESTLE) Approach for Risk Identification of the Tidal Industry in the United Kingdom”, Energies 2013, 6, 5023-5045.