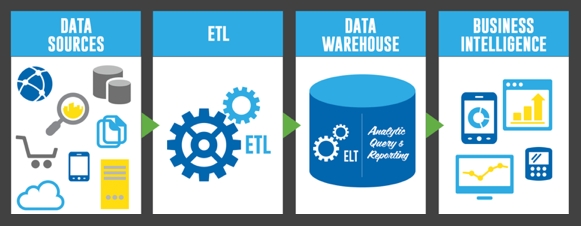
**A Project Report**

**On**

LEVERAGING CORPORATE STRATEGY USING

**DATA(ETL) TECHNIQUES**



*Submitted by*

**VIPUL GOYAL**

Trainee, ITS Department, EIL

**PREFACE**

**About This Project Report**

This report describes about the DATA (ETL) techniques and its use in detail for the corporate strategy which will be used for business opportunities and growth of the organization.

**About DATA (ETL)**

In computing, **Extract, Transform, Load** (**ETL**) refers to a process in database usage and especially in data warehousing. The ETL process became as Data extraction  is where data is extracted from homogeneous or heterogeneous data sources; data transformation where the data is transformed for storing in the proper format or structure for the purposes of querying and analysis; data loading  where the data is loaded into the final target database, more specifically, an operational data store, data mart or data warehouse.

Since the data extraction takes time, it is common to execute the three phases in parallel. While the data is being extracted, another transformation process executes while processing the data already received and prepares it for loading while the data loading begins without waiting for the completion of the previous phases.

ETL systems commonly integrate data from multiple applications (systems), typically developed and supported by different vendors or hosted on separate computer hardware. The disparate systems containing the original data are frequently managed and operated by different employees. For example, a cost accounting system may combine data from payroll, sales, and purchasing.

**BONAFIDE CERTIFICATE**

This is to certify that this project report entitled “**Leveraging The Corporate Strategy Using Data(ETL) Techniques**” submitted in ITS Department of Engineers India Limited is a bonafide record of work done by **VIPUL GOYAL** under my supervision from 5th June, 2017 to 28th July, 2017.

Place: New Delhi Mr. Niranjan Behera

Date: 28th July, 2017 (Coordinator)

ITS Department, EIL

**ACKNOWLEDGEMENT**

I would like to express my special thanks of gratitude to Mr. Gautam Bhalla (CGM-ITS) , Mr.Niranjan Behera (Coordinator) and Mr.Vinay Mangal (AGM) –Information Technology Services (ITS), department Engineers India Limited, New Delhi for providing me the opportunity to work in such an esteemed organization.

I would like to extend my indebtedness to them for guiding me in completing this project. I have benefited so much from the constructive criticism and suggestions given to me by them during the project.

I am extremely thankful to Engineers India Limited (EIL) for providing me with the opportunity to avail the excellent facilities and infrastructure in completing my vocational training.

VIPUL GOYAL

**TABLE OF CONTENTS**

**S.No. Title Page no.**

01PREFACE 02

02 BONAFIDE CERTIFICATE 03

03 ACKNOWLEDGEMENT 04

04 ABOUT THE ORGANISATION 07-13

4.1 EIL PROFILE

4.2 HISTORICAL BACKGROUND

4.3 SERVICES OFFERED BY EIL

05 ABOUT THE DEPARTMENT 14-15

5.1 INTRODUCTION

5.2 RANGE OF IT SERVICES

06 INTRODUCTION TO THE PROJECT 16-18

6.1 ABOUT ETL

07 TOOLS 19-24

7.1 SCRAPY

7.2 PYTHON(LANGUAGE)

7.3 ANACONDA(PYTHON DISTRIBUTION)

7.4 JUPYTER NOTEBOOK(FOR PYTHON)

7.5 JSON(FLAT FILE)

7.6 MONGODB

7.7 PHP(FOR DISPLAYING RESULT)

08 NLP(NATURAL LANGUAGE PROCESSING) 24-25

09 ONLINE REPOSITORIES 26

10 BATCH FILE 26

11 ENGINES AND FRAMEWORK USED WITH 27-29

PYTHON

11.1 SCRAPY

11.2 TWISTED

11.3 NLTK

11.4 PATHLIB

11.5 OS

11.6 JSON

11.7 FUZZYWUZZY

11.8 SYS

11.9 PYMONGO

12 METHODOLOGY 29-31

12.1 EXTRACTION OF DATA

12.2 PROCESSING OF DATA

12.3 STORAGE OF DATA

12.4 LOADING OF DATA

12.5 AUTOMATION

13 CURRENT USE 31

14 SCREENSHOTS (PROJECT) 32-36

14.1 ANACONDA PROMPT

14.2 OPENING JUPYTER NOTEBOOK

14.3 PROJECT DIRECTORY

14.4 RUNNING OF BATCH FILE

14.5 MLAB DATABASE BEFORE EXECUTION

14.6 CRAWLING WEBSITES

14.7 APPLYING NLP AND STORING IT ON

ONLINE DATABASE

14.8 DISPLAYING DATA

14.9 AUTOMATION

15 CONCLUSION 37

16 FURTHER SCOPE 37

17 BIBLIOGRAPHY 37

**4. ABOUT THE ORGANIZATION**

## *ENGINEERS INDIA LIMITED*

## *4.1 EIL Profile*

Engineers India Limited (EIL) was established on March 15, 1965, as a joint venture between the government of India and Bechtel Corporation of USA with the following prime objectives:

*“To establish, provide, maintain and perform engineering and related technical and consultancy services for petroleum projects including but not limited to, petroleum refineries, oil field development, oil and gas pipelines, petrochemical facilities, chemical intermediaries and all other types of industrial projects”*

In addition to petroleum refineries, EIL has diversified into other fields such as pipelines, chemicals and oil and gas processing, etc. EIL provides a complete range of project services in the fields and has emerged as South Asia’s leading design and engineering company.

EIL has under gone through more than 4000 assignments, 250 projects worth more than US$ 35 billion in installed cost, successfully completed and operating smoothly creating an array of satisfied clients and rising the turnover and profits. The major projects under taken by the company includes 30 Petroleum Projects, 6 Petroleum Complexes, 200 Offshore Platforms, 31 Oil and Gas Processing Projects, 25 Mining and Metallurgical Projects, 8 Fertilizers Projects, 31 Pipelines Projects and 11 Ports and Terminals. EIL has also worked abroad creditability in several countries of West Asia, North Africa, Europe and South East Asia including Algeria, Abu Dhabi, Iran, Kuwait, Malaysia, Norway, Qatar, Saudi Arabia, Sri Lanka, UAE and Vietnam.

### 4.1.1 Vision Statement

*“To become a globally Competitive* *EPC & Consultancy Organization”*

### 4.1.2 Mission Statement

* To achieve international standards of excellence with a focus on customer satisfaction.
* To provide high quality, safe and energy-efficient services in Process Design, Engineering, Procurement, Construction with overall project Management in Hydrocarbon, Metallurgy, Power and Ports & Terminals as well as Information Technology and other selected sectors of Industry.
* To achieve prominence in developing, adopting and assimilating state-of-the-art technology for competitive advantage.
* To maximize creation of wealth, value and satisfaction for stake holders.
* To foster a culture of participation and innovation for employees' growth and contribution through a climate of fairness and transparency in operation.
* To acquire and provide technology and services through sustained Research & Development and linkage among Engineering Organizations, Equipment Manufacturers, Operating Companies and R&D Organizations to upgrade technologies on a continuous basis.
* To cultivate high standards of ethics and quality for a strong corporate identity and brand equity.
* To help enrich the quality of life of the community and preserve ecological balance and heritage through the services provided with a strong environmental conscience.

## *4.2 Historical Background*

The foundation of the consultancy profession in India was laid soon after independence. The sixties saw the advent of major consultancy and engineering organization in the country with the setting up of a number of design and engineering organizations, in both private sector as well as public sector. Of these, Engineers India Ltd. Was the most prominent which was created in 1965 as a joint venture of the government of India (holding 51% shares) and Bechtel, a U.S. based company (holding 49% shares) to provide engineering and related services for petroleum and other industrial projects. In 1967, EIL became a wholly owned Government of India Undertaking.

Engineers India Limited has been serving the process industry providing a complete range of project services. Initially, it started with engineering work for petroleum refineries and diversified over the years, to add other fields of activity as well as other services. In response to the changing business environment, Engineers India went through several stages of development and has gradually evolved into an engineering company serving a wide range of industries including petroleum refineries, petrochemicals, oils and gas processing projects, pipelines, offshore platforms, fertilizers, chemical fibers and metallurgical industries providing a complete range of project services .

It provides consultancy in the following fields of operation:

* Petroleum
* Pipelines
* Oil and Gas Processing
* Fertilizers
* Power Ports & Terminals
* Petrochemicals
* Chemicals
* Metallurgy

### 4.2.1 Petroleum Refineries

Engineers India Limited has provided its services for over a dozen projects with a combined refining capacity of 23 million tones/annum (460,000 bbls/day) and is working on several other projects with a total refining capacity of over 25 million tones/annum (500,000 bbls/day). The projects include grassroots as well as expansion revamp projects of all refining companies in India apart from refinery projects abroad. EIL has developed skills to such an extent that barring the process design of a few licensed units; EIL can execute complete petroleum refinery projects on its own. In addition to technologies for the main refinery units, EIL also has technologies for lube refinery complexes.

### 4.2.2 Fertilizers

Engineers India Limited provides complete services starting with feasibility study up to commissioning for fertilizer plants based on gas, naphtha and fuel oil. EIL has worked with renowned licensors/contractors on various assignments for ammonia, urea and phosphates fertilizer plants, in India and abroad.

### 4.2.3 Pipelines

Engineers India has the capability to plan and execute long distance cross country and submarine pipelines for transportation of crude oil, petroleum products, gases, two phase fluids and slurries. It provides services such as conceptual scheme, feasibility study, on site investigations and route survey, cathodic protection and telemetry, telecommunication and tele supervisory control for pipeline projects besides other project services such as design and engineering, procurement, construction supervision and project management.

### 4.2.4 Oil and Gas processing

Engineers India Limited has expertise for Crude Oil processing plants such as:

* Crude Gathering Stations
* Multistage Separation
* Crude Desalting
* Dehydration and Heavy Oil/Sour Oil Processing.

It also has capabilities for implementing Gas Conditioning and Processing Plants including Gas Collection Systems, Compressor Stations, and Solid/Liquid Desiccant Dehydration, Dew point Depression, Impurities and Acid gases removal and Cryogenic Natural Gas Liquids recovery such as LPG and Ethane / Propane recovery. It has handled a number of such plants for the Oil and Natural Gas Corporation Limited and Gas Authority of India Ltd. It has also provided consultancy for LPG and LNG projects of Sonatrach in Algeria.

### 4.2.5 Power

In the field of power, EIL has undertaken assignments for captive power plants in large process plants such as petroleum refineries, petrochemical complexes, etc. and has undertaken feasibility studies for refinery residue based power projects.

EIL has association with: -Foster Wheeler Italiana for Residue Based Power Plants Kema, Netherlands for Renovation & Modernization and Life extension studies for Existing Power Plants.

### 4.2.6 Ports and terminals

The Ports & Terminals Division of EIL takes up projects such as jetties, SPMs, MBMs at waterfronts or near to shore and related submarine pipelines with onshore facilities like terminals, etc. The areas of operation include:

* Master Plans for Ports
* Technical assistance in Marine Surveys
* Port Facilities/Jetties
* Inland Terminals for receipt, storage and dispatch of liquid cargo in both pressurized and cryogenic conditions

Offshore Handling Facilities including:

* Single Buoy Moorings
* Multi Buoy Moorings
* Submarine Pipelines

Port based Terminals for receipt, storage and dispatch of:

* Hydrocarbon products
* Chemicals & Petrochemicals
* Dry cargo both in bulk and bagged condition

### 4.2.7 Ocean Engineering

Engineers India's Ocean Engineering Division was formed in 1971 and has been involved in the development of India's offshore oil and gas fields since then. Engineers India has wide capabilities for providing full range of services needed to plan, design, engineer, construct and commission ocean engineering projects such as oil and gas platforms, offshore oil terminals, oil and fertilizer loading berths, intake structures, SBMs, offshore lighthouses, floating pump houses and fabrication yards.

4.2.8 Chemical

EIL has expertise for chemical plants such as membrane cell technology based caustic soda plants, soda ash, vaccine, insecticides etc., for which complete range of services can be offered.

### 4.2.9 Metallurgy

In the field of metallurgy, EIL can render a comprehensive range of consultancy and engineering services for zinc, lead, aluminum, copper, cadmium, silver, nickel, magnesium, titanium, etc. and also for sponge iron. In addition, EIL can provide services for mine development and beneficiation plants. The specialized services include Studies for Scaling up of Pilot Plant to Industrial Plant, Process & Licensor selection, Optimization of the process parameters, Energy Conservation etc. The areas of operation encompass:

* Exploration data review, Geo statistical Ore reserve estimation and Ore Body modeling.
* Planning and Design of Open Pit Mines, Pit Optimization Studies and Quality control.
* Environmental studies for Mines.
* Ore handling, storage and long distance conveying.
* Design of Mineral Processing Plant and simulation & Optimization of crushing, grinding and beneficiation circuits including conventional as well as column flotation cells.
* Hydrometallurgical plants such as alumina refineries and lead smelters.
* Pyro/electro metallurgical plants such as primary Aluminum and lead smelters. Design of environment friendly tailings disposal/ stacking systems.

## *Services Offered by EIL*

### 4.3.1 Feasibility Studies

EIL's experience in a variety of fields is available to clients for expansion or diversification of their activities. Services offered include preliminary studies to examine the viability of project, market potential of products, techno-economic evaluation, capital and operating cost estimates etc

### 4.3.2 Pre-feasibility Studies

To examine the viability of the project, market potential of products,techno-economic evaluation, capital & operating cost estimates etc.

### 4.3.3 Detailed Feasibility Studies

EIL also undertakes detailed feasibility studies to form the final cost of the project. These detailed feasibility reports are also bankable to ensure that financial institutions can use them for lending purposes.

EIL's project management services comprise direction and coordination of the activities leading to the completion of a job in accordance with terms of the contract, in compliance with the job schedule and to the satisfaction of the client. The efforts of various agencies such as licensors, sub-contractors & EIL's different departments are brought together in a logical pattern to result in fruitful culmination of the project.

### 4.3.4 Heat and Mass Transfer Equipment Design

Heat and Mass Transfer Division (HMTD) specializes in design and revamp of heat & mass transfer equipment & related process package systems as part of overall EIL project or as individual direct assignment to meet the requirement of new as well as existing plants. Experience of HMTD in design and revamp of heat and mass transfer equipment system is unparalleled.

The experienced manpower resources in HMTD carry out a wide range of services for the following heat & mass transfer equipment, process and energy systems:

* Fired Heaters & Air Preheating System
* Heat Exchangers
* Trays
* Packing & Packed Tower Internals
* Gas Dehydration System and Oil - Water - Gas Separators
* Vacuum Ejector System, Process Waste Heat & CO Boiler and Incinerators
* Deaerators
* De salters

4.3.5 Detailed Engineering

The process design package is converted into detailed engineering drawings and documents by specialist civil, structural, mechanical, and electrical & instrumentation engineers, Computer aided design techniques are extensively used as design aides to help develop optimum layout & design to assure maximum construction, operational & maintenance convenience.

Detailed engineering leads to identification & specification of equipment and materials for procurement & needs of construction activities at site.

Specialized Services are provided in the fields of static & dynamic analysis of equipment, structures, geotechnical engineering, rotating equipment & troubleshooting.

### 4.3.6 Project Management Consultancy

The experience gained in the implementation of complex integrated projects has enabled Engineers India to undertake the role of Project Management Consultant. As Project Management Consultant, EIL provides management services required for smooth implementation of a complex project. It takes full responsibility for the overall co-ordination and integration of functioning of all agencies engaged in the implementation of the entire project and ensures the completion on schedule and within the budgeted cost.

EIL's scope as Project Management Consultant broadly covers the following:

* Conceptualization of the Project
* Technology Selection.
* Developing Design Basis.
* Selection of EPC Contractors.
* Project Management including supervision, co-ordination, planning and scheduling services, etc.
* Quality Assurance at all stages of the project implementation, viz. process design, detailed engineering, procurement and construction.
* Assistance in inspection, testing, pre-commissioning and commissioning of facilities.

### 4.3.7 Environmental Services

EIL offers extensive services in the field of environmental engineering such as Environment impact assessment, Waste water treatment, Water treatment, Solids and hazardous waste management, Environmental audits, Resource recovery and recycle, Pollution prevention and Waste minimization, etc. These services are provided for Petroleum Refineries, Petrochemical Plants, Pipeline Projects, Oil/Gas Installations and Chemical Ports, Terminals, SBM's and Jetties and Non-ferrous Metallurgical projects. EIL's close familiarity with the technologies involved in the process industry gives it a unique advantage for undertaking environment-related assignments.

### 

### 4.3.8 Planning and Scheduling

EIL ensures effective and timely execution of projects through detailed planning and micro scheduling along with continuous monitoring. For this purpose, the industry standard PRIMA VERA software is used for computerized network analysis (PERT/CPM). Other software packages such as HOST, CPS, MCS, and COSMAS etc. are used for home office scheduling, construction, and planning and material control. Detailed schedules for engineering, ordering, manufacturing, delivery, tendering and construction are developed. These schedules are reviewed and updated and project completion outlook analyzed and corrective actions taken regularly. In addition, physical progress of the project is measured and reported through a well-designed reporting system, which enables complete evaluation of project performance.

### 4.3.9 Construction Management

EIL's Construction Division provides total construction management services at site including selection of construction contractors, warehouse management, quality control/quality assurance, process monitoring & scheduling, review of heavy erection schemes, safety etc. Depending on client's need, EIL takes total site responsibility from survey & soil investigation to mechanical completion & assistance in commissioning. Use of modern construction techniques, innovative construction procedures, emphasis on QA/QC are some of the hallmarks based on which it has been possible for EIL to reduce construction time of projects yet providing work of highest quality.

### 4.3.10 Cost Engineering

EIL’s cost engineering services help the project management team in exercising practically online control of the costs vis-à-vis project execution budgets through application of latest techniques of Project Cost Management. Large dynamic database & in-house developed programs operating on PC platform are used in estimating costs of work packages prior to their award. A periodic 'Cost-to-Completion' projection is made through in-house developed program "CEMOS" software to assist the Project Manager & Owner in taking timely decisions & actions. The cost engineering systems employed are kept updated through the latest developments in cost management techniques as also through experience gained while working on international projects.

### 4.3.11 Procurement

EIL offers comprehensive procurement services through a highly professional purchase team, effective monitoring and expediting group, a well experienced and qualified multi-disciplinary inspection force and a specialist group for route survey and planning for multi model movement of over dimensional consignments.

Following comprehensive procurement services are provided

* Vendor identification, evaluation & enlistment.
* Entire purchase activity, floating of enquiry to placement of order.
* Order placement
* Progress monitoring & Expediting
* Inspection
* Transportation planning & control
* Safety certification under Static & Mobile Pressure Vessels (SMPV) rules of the Chief Controller of Explosives
* Statutory inspection in Europe & Japan under Indian Bouler Regulations.

A databank, providing significant information on capable vendors all over the world, is maintained and constantly updated. EIL's twelve inspection offices in the country provide competitive and prompt services in India. Overseas offices in Tokyo & London offer procurement/inspection services outside India in the Far East & Europe.

### 4.3.12 Commissioning and Plant Start up

EIL's Commissioning team associates itself with the project right from the process design stage. It reviews process & engineering design documents especially with respect to operability & safety. The team also prepares operation manuals providing instructions for plant start-up, shutdown & handling various emergencies. It provides guidelines and supervises pre commissioning & commissioning activities at site. EIL's responsibility continues till production of specification products has been established to the full satisfaction of the customers. The team assists process department in conducting performance guarantee tests in order to establish plant operation at design throughputs with product of specified quality. EIL also provides technical guidance to customers in the area of troubleshooting.

**5. ABOUT THE DEPARTMENT**

## *5.1 Introduction*

Information Technology Services (ITS), a division of EIL has a well-knit team of over 100 multidisciplinary professionals who are well qualified and experienced in providing software related and networking design services. Over the last 25 years, ITS has developed a large number of integrated software packages for use within EIL in the areas of process design, engineering, procurement, project management and construction. This experience along with the domain knowledge in various functional disciplines has given an edge to ITS for providing value added services to its clients in Oil Industry in particular and others in general. ITS has spearheaded the task of making EIL a leader both in terms of software usage and having state-of-the-art computing facilities and communication infrastructure leading to enterprise connectivity.

ITS over the years has executed a number of offshore and onshore software development assignments for clients in India and abroad. The application areas covered are engineering solutions, business applications (viz. materials management, maintenance management, financial accounting, and personnel management), integrated plant information systems, intranet development, operations research etc.

An overview of range of services along with assignments handled and clientele highlighting EIL's IT experience is enumerated hereafter.

### 

## *5.2 Range of IT Services*

### Preparation of a Master Plan and Its Implementation

EIL has executed a number of assignments in the areas of configuration assessment, evaluation, selection and acceptance testing of computer and communication systems and implementation for clients like:

* Gas Authority of India Ltd.
* Bongaigaon Refinery & Petrochemicals Ltd.
* Electricity Boards etc.

EIL has also executed studies involving preparation of Refinery wide Master Plans taking into consideration the integration of DCS, process functions & management information systems for clients like:

* Indian Oil Corporation,
* Cochin Refineries
* Bharat Petroleum Corporation Limited.

### Customized Software Development for Engineering and Business Application

EIL has undertaken a number of customized software development and implementation assignments in the following application areas:

* Implementation of Plant Design Systems on PDS & PDMS platforms for Aker Corporation, Norway & PIDEC in Iran.
* On-line Integrated Materials Information System covering activities like indents, purchase, stores issues/receipts, material accounting for operating plants namely Indian Petrochemicals Corporation Ltd., Oil & Natural Gas Corporation, Bongaigaon Refinery & Petrochemicals Ltd.
* Computer Aided Maintenance Information System covering preventive & predictive maintenance, work order, maintenance cost & breakdown history for BIPC, Iran. Financial Management Systems including books, ledgers (main & subsidiary), balance sheet, & asset account for major electrical utilities.

### Plant Database Creation & Document Management

EIL has developed software, which provides for value added transfer of integrated plant information system from EPC contractor to the client on electronic media. This software has been implemented for Panipat Refinery Project and many other jobs are in advanced stage of finalization.

### Internet Application Development and Implementation

EIL has developed expertise in development of JAVA & HTML based applications for implementation on Intranet. In addition to having Intranet for in-house needs, an assignment has recently been completed for Gas Authority of India Ltd.

### Enterprise Recourse Planning (ERP) Implementation

EIL has trained its software specialists and domain experts on the implementation of ERP solutions. Currently a number of jobs are planned to be taken for Oil Industry in India in association with our business partners M/s RAMCO for their ERP product, namely MARSHAL.

### Operations Research (Linear Programming, Transport Models and Queuing)

EIL has executed a number of assignments involving OR techniques for the Ministry of Petroleum, Govt. of India and also a study based on queuing model for a port in Malaysia. Major investment decisions have been taken by the clients based on these optimization studies.

**6. INTRODUCTION TO THE PROJECT**

Corporate Strategy is an important aspect for the growth and sustainability of the organization. The growth potential, functional diversification , competition in the environment of the organization are suitably addressed with corporate strategy in the business for short term as well as long term. The strategy can be identified and analyzed to meet the need of the organization from time to time. The growth in information technology is used to leverage the corporate strategy.

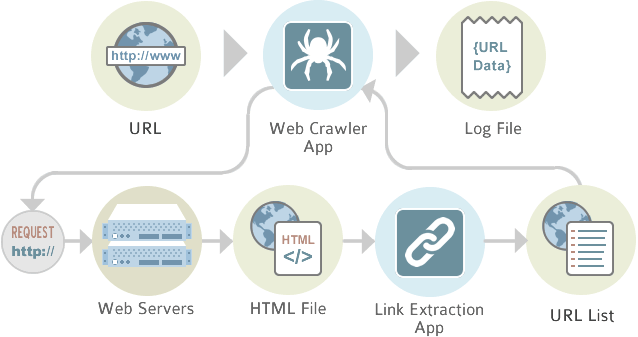
It is extremely difficult to achieve the objective due to the dynamic nature of enterprise, availability of data, heterogeneous systems and requirement of the top management.

The IT tools and techniques are used to identify, collect, summarize, analyze the data, information in various forms towards the corporate strategy. The ETL, is one of the tools and techniques widely used for strategy and business purposes which is an effective tool for the decision makers. It takes data/information from various sources, transforms into specific need purpose and loads into the decision makers. It is carried out without any delay which, is a critical factor in any decision making process.

The details regarding ETL and methodologies are described in detail in this report.

**6.1 About DATA (ETL)**

**6.1.1 Overview**



**FIGURE DEPICTING DATA EXTRACTION**

ETL - extract, transform and load - is the set of processes by which data is extracted from numerous databases, applications and systems, transformed as appropriate, and loaded into target systems - including, but not limited to, data warehouses, data marts, analytical applications, etc.  
The first part of the extract, transform and load (ETL) process is understanding the data sources. The transformations are organization-specific and Integration is sometimes included in the ETL process; because it requires an in-depth knowledge of the organization and its business.

**Data Modeling.** While there are modeling commonalities within a vertical industry, every organization has its own way of doing business; these unique processes should be included in the models  
  
**Data warehouse Design.** Ignify provides datawarehouse design services with consultants that understand the processing requirements and have the ability to deliver high performance data warehouses.

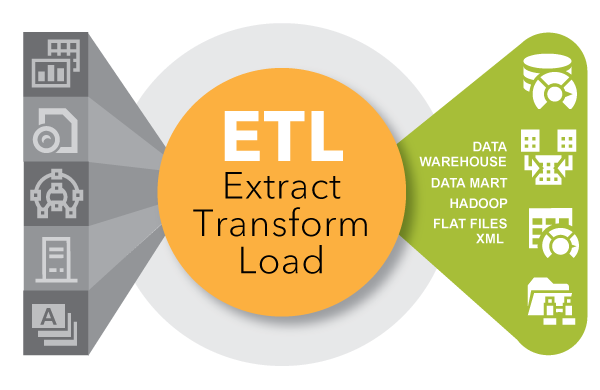
**6.1.2 History**

ETL gained popularity in the 1970s when organizations began using multiple data repositories, or databases, to store different types of business information. The need to integrate data that was spread across these databases grew quickly. ETL became the standard method for taking data from disparate sources and transforming it before loading it to a target source, or destination.

In the late 1980s and early 1990s, data warehouses came onto the scene. A distinct type of database, data warehouses provided integrated access to data from multiple systems – mainframe computers, minicomputers, personal computers and spreadsheets. But different departments often chose different ETL tools to use with different data warehouses. Coupled with mergers and acquisitions, many organizations wound up with several different ETL solutions that were not integrated.

**6.1.3 Description**

Businesses have relied on the ETL process for many years to get a consolidated view of the data that drives better business decisions. Today, this method of integrating data from multiple systems and sources is still a core component of an organization’s data integration toolbox.



ETL is used to move and transform data from many different sources and load it into various targets, like Hadoop.

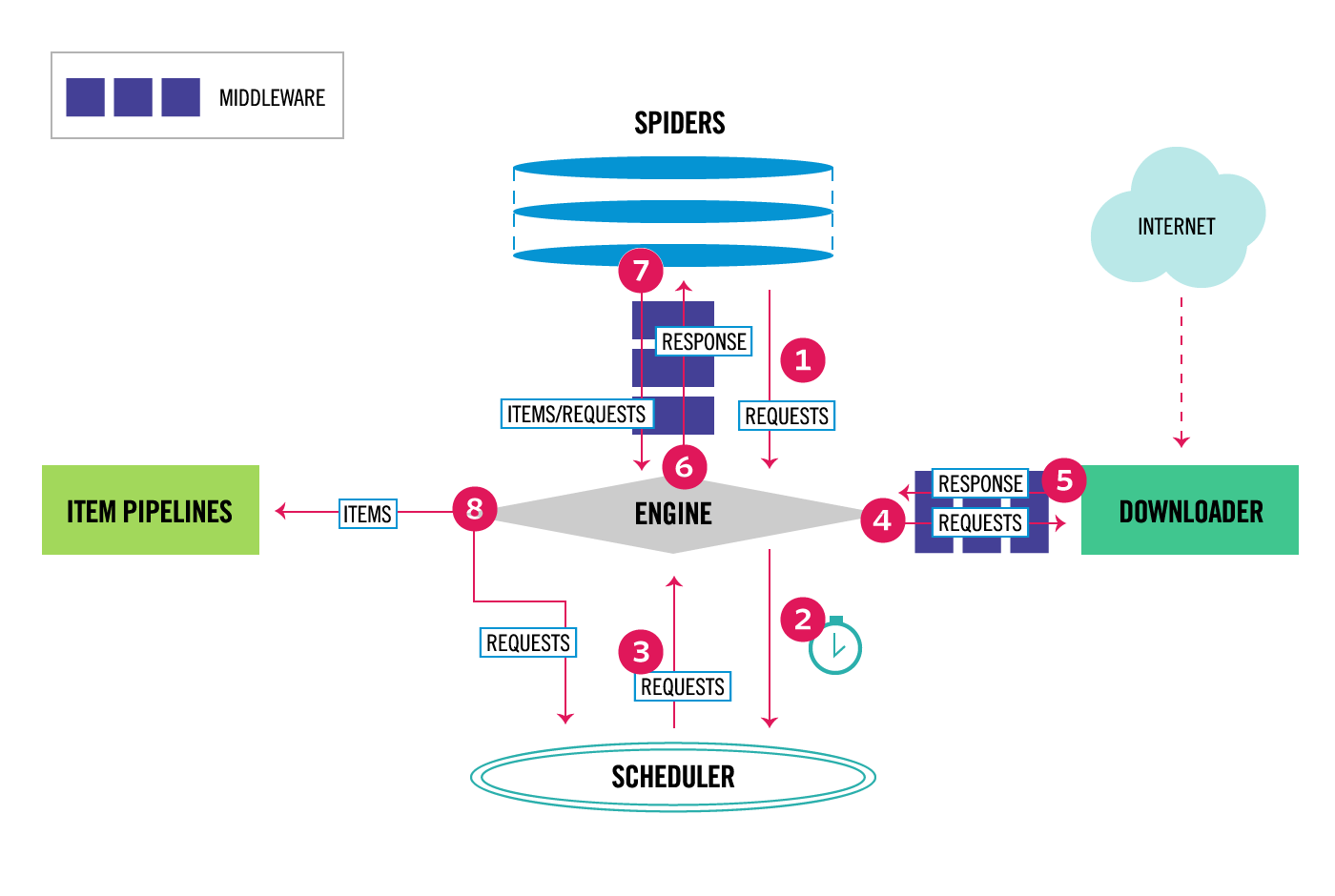
* When used with an enterprise data warehouse (data at rest), ETL provides deep historical context for the business.
* By providing a consolidated view, ETL makes it easier for business users to analyze and report on data relevant to their initiatives.
* ETL can improve data professionals’ productivity because it codifies and reuses processes that move data without requiring technical skills to write code or scripts.
* ETL has evolved over time to support emerging integration requirements for things like streaming data.
* Organizations need both ETL and ETL to bring data together, maintain accuracy and provide the auditing typically required for data warehousing, reporting and analytics.

**7. TOOLS**

**7.1 SCRAPY**

An open source and collaborative framework for extracting the data you need from websites. In a fast, simple, yet extensible way. That means you can use it to manage requests, preserve user sessions, follow redirects, and handle output pipelines. It also means you can swap out individual modules with other Python web scraping libraries.

Scrapy also provides reusability of crawler by scaling it and to manage complex data pipelines, or cook up some other sophisticated spider.



**7.2 PYTHON (LANGUAGE)**

Python is a widely used high-level programming language for general-purpose programming. An interpreted language, Python has a design philosophy which emphasizes code readability (notably using whitespace indentation to delimit code blocks rather than curly brackets or keywords), and a syntax which allows programmers to express concepts in fewer lines of code than might be used in languages such as C++ or Java. The language provides constructs intended to enable writing clear programs on both a small and large scale.

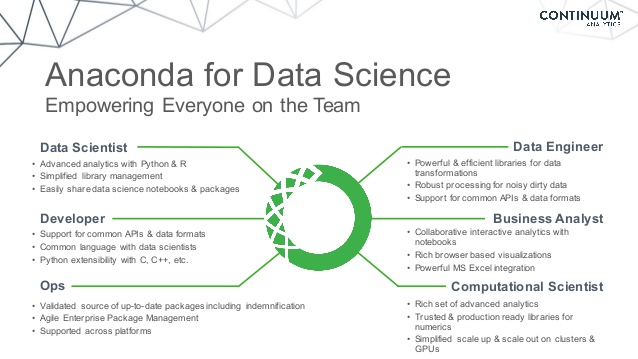
Python features a dynamic type system and automatic memory management and supports multiple programming paradigms, including object-oriented, imperative, functional programming, and procedural styles. It has a large and comprehensive standard library.

Python interpreters are available for many operating systems, allowing Python code to run on a wide variety of systems.

**Python Version Used in the Project: 3.6.1**

**7.3 ANACONDA(Python Distribution)**

**Anaconda** is a freemium open source distribution of the Python and R programming languages for large-scale data processing, predictive analytics, and scientific computing, that aims to simplify package management and deployment. Package versions are managed by the package management system *conda*.



**7.4 JUPYTER NOTEBOOK(FOR PYTHON)**

#### The Jupyter Notebook is an open-source web application that allows us to create and share documents that contain live code, equations, visualizations and explanatory text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, machine learning and much more.

**7.4.1 Features**

**7.4.1.1 Languages of choice**

The Notebook has support for over 40 programming languages, including those popular in Data Science such as Python, R, Julia and Scala.

**7.4.1.2 Share Notebooks**

Notebooks can be shared with others using email, Dropbox, GitHub and the Jupyter Notebook Viewer.

**7.4.1.3 Interactive widgets**

Code can produce rich output such as images, videos, LaTeX, and JavaScript. Interactive widgets can be used to manipulate and visualize data in Realtime.

**7.4.1.4 Big data integration**

Leverage big data tools, such as Apache Spark, from Python, R and Scala. Explore that same data with pandas, scikit-learn, ggplot2, dplyr, etc.

**7.5 JSON(FLAT FILE)**

**7.5.1 Flat File**

A flat file contains records that have no structured interrelationship. A flat file typically consists of a text file, from which all word processing or other structure characters or markup have been removed.

**7.5.2 JSON**

**JSON** (JavaScript Object Notation) is a lightweight data-interchange format. It is easy for humans to read and write. It is easy for machines to parse and generate. It is based on a subset of the JavaScript Programming Language, Standard ECMA-262 3rd Edition - December 1999. JSON is a text format that is completely language independent but uses conventions that are familiar to programmers of the C-family of languages, including C, C++, C#, Java, JavaScript, Perl, Python and many others. These properties make JSON an ideal data-interchange language.

**7.6 MONGODB**

MongoDB is a free and open-source cross-platform document-oriented database program. Classified as a NoSQL database program, MongoDB uses JSON-like documents with schemas. MongoDB is developed by MongoDB Inc. and is free and open-source, published under a combination of the GNU Affero General Public License and the Apache License.

**7.6.1 Main Features**

**7.6.1.1 Ad hoc queries**

MongoDB supports field, range queries, regular expression searches. Queries can return specific fields of documents and also include user-defined JavaScript functions. Queries can also be configured to return a random sample of results of a given size.

**7.6.1.2 Indexing**

Fields in a MongoDB document can be indexed with primary and secondary indices.

**7.6.1.3 Replication**

MongoDB provides high availability with replica sets. A replica set consists of two or more copies of the data. Each replica set member may act in the role of primary or secondary replica at any time. All writes and reads are done on the primary replica by default. Secondary replicas maintain a copy of the data of the primary using built-in replication. When a primary replica fails, the replica set automatically conducts an election process to determine which secondary should become the primary. Secondaries can optionally serve read operations, but that data is only eventually consistent by default.

**7.6.1.4 Load balancing**

MongoDB scales horizontally using sharding. The user chooses a shard key, which determines how the data in a collection will be distributed. The data is split into ranges (based on the shard key) and distributed across multiple shards. (A shard is a master with one or more slaves.). Alternatively, the shard key can be hashed to map to a shard – enabling an even data distribution.

MongoDB can run over multiple servers, balancing the load or duplicating data to keep the system up and running in case of hardware failure.

**7.6.1.5 File storage**

MongoDB can be used as a file system with load balancing and data replication features over multiple machines for storing files.

This function, called Grid File System, is included with MongoDB drivers. MongoDB exposes functions for file manipulation and content to developers. GridFS is used in plugins for NGINX and lighttpd. GridFS divides a file into parts, or chunks, and stores each of those chunks as a separate document.

**7.6.1.6 Aggregation**

MapReduce can be used for batch processing of data and aggregation operations.

The aggregation framework enables users to obtain the kind of results for which the SQL GROUP BY clause is used. Aggregation operators can be strung together to form a pipeline – analogous to Unix pipes. The aggregation framework includes the $lookup operator which can join documents from multiple documents, as well as statistical operators such as standard deviation.

**7.6.1.7 Server-side JavaScript execution**

JavaScript can be used in queries, aggregation functions (such as MapReduce), and sent directly to the database to be executed.

**7.6.1.8 Capped collections**

MongoDB supports fixed-size collections called capped collections. This type of collection maintains insertion order and, once the specified size has been reached, behaves like a circular queue.

**7.7 PHP (FOR DISPLAYING RESULT)**

PHP is a server scripting language, and a powerful tool for making dynamic and interactive Web pages. PHP is a widely-used, free, and efficient alternative to competitors such as Microsoft's ASP.

**8. NLP(NATURAL LANGUAGE PROCESSING)**

* Natural language processing (NLP) is a field of computer science, artificial intelligence and computational linguistics concerned with the interactions between computers and human (natural) languages, and, in particular, concerned with programming computers to fruitfully process large natural language corpora. Challenges in natural language processing frequently involve natural language understanding, natural language generation (frequently from formal, machine-readable logical forms), connecting language and machine perception, dialog systems, or some combination thereof.
* Features of Natural Language Processing used in the project are:
  + Tokenization
  + Stop Word Removal
  + Stemming
  + Lemmatizing

**8.1 TOKENIZATION**

Tokenization is the process of demarcating and possibly classifying sections of a string of input characters. The resulting tokens are then passed on to some other form of processing. The process can be considered a sub-task of parsing input.

**8.2 STOP WORD REMOVAL**

The process of converting data to something a computer can understand is referred to as **pre-processing.**One of the major forms of pre-processing is to filter out useless data. In natural language processing, useless words (data), are referred to as stop words.

**8.2.1 STOP WORDS**

A stop word is a commonly used word (such as “the”, “a”, “an”, “in”) that a search engine has been programmed to ignore, both when indexing entries for searching and when retrieving them as the result of a search query.

**8.3 STEMMING**

Stemming is the process of reducing inflected (or sometimes derived) words to their word stem, base or root form—generally a written word form. The stem need not be identical to the morphological root of the word; it is usually sufficient that related words map to the same stem, even if this stem is not in itself a valid root.

A stemmer for English, for example, should identify the string "cats" (and possibly "catlike", "catty" etc.) as based on the root "cat", and "stems", "stemmer", "stemming", "stemmed" as based on "stem". A stemming algorithm reduces the words "fishing", "fished", and "fisher" to the root word, "fish". On the other hand, "argue", "argued", "argues", "arguing", and "argus" reduce to the stem "argu" (illustrating the case where the stem is not itself a word or root) but "argument" and "arguments" reduce to the stem "argument".

**8.4 LEMMATIZATION**

Lemmatization in linguistics is the process of grouping together the inflected forms of a word so they can be analysed as a single item, identified by the word's lemma, or dictionary form.

In computational linguistics, lemmatization is the algorithmic process of determining the lemma of a word based on its intended meaning. Unlike stemming, lemmatization depends on correctly identifying the intended part of speech and meaning of a word in a sentence, as well as within the larger context surrounding that sentence, such as neighboring sentences or even an entire document. As a result, developing efficient lemmatization algorithms is an open area of research.

For example the word "better" has "good" as its lemma. This link is missed by stemming, as it requires a dictionary look-up.

**9. ONLINE REPOSITORIES**

**9.1 mLAB**

mLab is a fully managed cloud database service that hosts MongoDB databases. mLab runs on cloud providers Amazon, Google, and Microsoft Azure, and has partnered with platform-as-a-service providers.

**9.1.1 MongoDB-as-a-Service**

### 9.1.1.1 Infrastructure-as-a-Service (IaaS) Partners

* Amazon Web Services (AWS)
* Google Cloud Platform
* Microsoft Azure

### 9.1.1.2 Platform-as-a-Service (PaaS) Partners

* Heroku

**10. BATCH FILE AUTOMATION**

Making use of batch files and scripts in a computer network is an essential part of a network administrator's tool kit. There are many tasks/functions that, when automated, improve systems management capabilities and reliability. An automated task can be run at any time of day. Some tasks that can be automated are:

* System drive defragmentation
* Permission set/update on home directories
* Housekeeping (removing temporary files from servers or home directories)
* Time synchronization
* Data file backup
* Log file export or backup
* Restarting system services
* Checking system up time

**11.ENGINES AND FRAMEWORK USED WITH PYTHON**

**11.1 SCRAPY**

An open source and collaborative framework for extracting the data you need from websites.

**11.2 TWISTED**

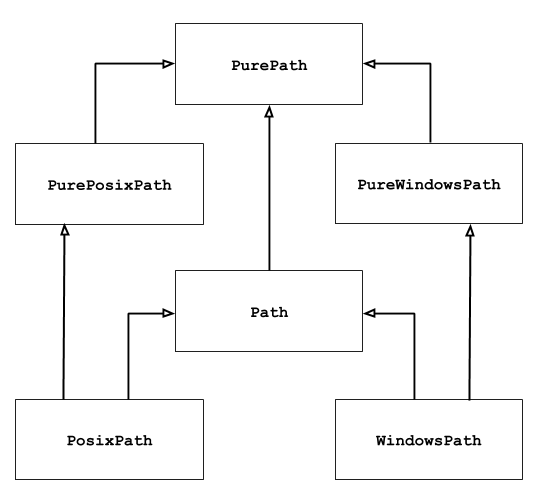
Twisted is an event-driven networking engine written in Python and licensed under the open source ​MIT license. Twisted runs on Python 2 and an ever growing subset also works with Python 3

**11.3 NLTK**

NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum.

**11.4 PATHLIB**

This module offers classes representing filesystem paths with semantics appropriate for different operating systems. Path classes are divided between pure paths, which provide purely computational operations without I/O, and concrete paths, which inherit from pure paths but also provide I/O operations.



**11.5 OS**

This module provides a portable way of using operating system dependent functionality. If you just want to read or write a file see open(), if you want to manipulate paths, see the os.path module, and if you want to read all the lines in all the files on the command line see the fileinput module. For creating temporary files and directories see the tempfile module, and for high-level file and directory handling see the shutil module.

**11.6 JSON**

The [json](https://docs.python.org/2/library/json.html) library can parse JSON from strings or files. The library parses JSON into a Python dictionary or list. It can also convert Python dictionaries or lists into JSON strings.

**11.7 FUZZYWUZZY**

FuzzyWuzzy has been developed and open-sourced by SeatGeek, a service to find sport and concert tickets.

Fuzzy string matching like a boss. It uses Levenshtein Distance to calculate the differences between sequences in a simple-to-use package.

**11.8 SYS**

This module provides access to some variables used or maintained by the interpreter and to functions that interact strongly with the interpreter. It is always available.

**11.9 PYMONGO**

Python driver for MongoDB

The PyMongo distribution contains tools for interacting with MongoDB database from Python. The bson package is an implementation of the BSON format for Python. The pymongo package is a native Python driver for MongoDB. The gridfs package is a gridfs implementation on top of pymongo.

**12. METHODOLOGY**

## 12.1 Extraction of Data

### All the accessible data of 5 websites has been extracted and stored in a JSON File in form of a list.

### For extraction of data one spider is attached to every website and then all these spiders are attached to a crawler that crawls and access the information of websites.

### I used Jupyter notebook as a platform using Anaconda prompt command.

**12.2 Processing of Data**

**12.2.1 Removal of Data without specified keywords**

* The content from JSON File (which was created after data extraction) are fetched.
* All the list in the content are tokenized and the list which do not contain specified keywords are removed from the file.

**12.2.2 Removal of character literals**

* From the remaining list all the character literals are replaced by an empty string.

**12.2.3 Removal of redundant data**

**12.2.3.1 Removal of stop words**

* Stop words from all the list in the left over content are removed

**12.2.3.2 Filtering of data**

* Now the data present in the list are filtered using the process of stemming and lemmatizing .
* This is a step which brings all the list present in the JSON File to basic language structure.

**12.2.3.3 Checking Similar List**

* Now all the list are compared to one another to find similar list pair.
* One of the List from each Similar List Pair is removed from the JSON File.
* Now the JSON file with all the relevant data is obtained.

**12.3 STORAGE OF THE DATA**

* Considering the minimum storage usage, the online repository mLAB providing assistance for Mongo DB is used.
* The data is stored and retrieved from database through a script written in python which uses the MongoDB standard URI to create connection.
* In order to obtain the minimum storage the previous data is removed automatically and thus refreshed by a new database.
* The flat file used is JSON file which is thus stored in online database in form of documents with unique autogenerated ID’s.

**12.4 LOADING OF THE DATA**

* The data is loaded on a simple web page designed using PHP attached with a login page to maintain data integrity.
  1. **AUTOMATION**

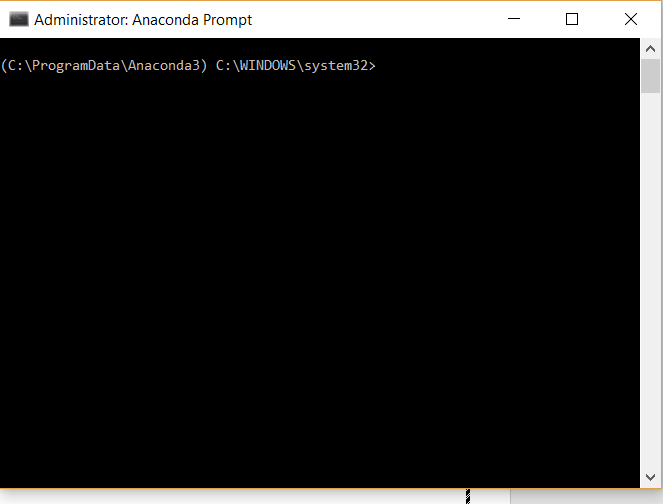
* All the above task are bundled into a single batch file, which runs after every 15 minutes so the data extracted will refresh continuously after time interval of every 15 minutes, hence providing the latest data from the websites to the user.

**13.CURRENT USE**

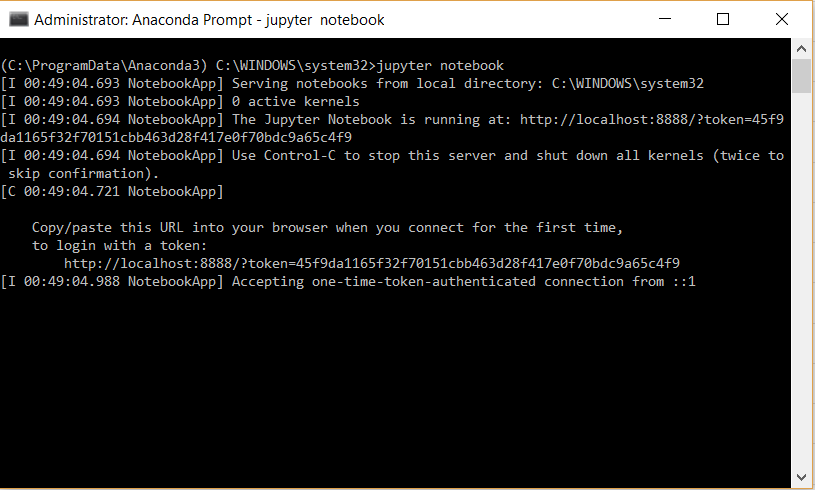
* My project provides news related to many fields of petroleum, which in turn can be used by the organization to gain knowledge about the latest deal,bids,mining and other aspects related to petroleum that are going on around the world
* My project will keep the organization upto date which will not only save there time but will help them to plan more systematically and efficiently.

**14. SCREENSHOTS**

**14.1 ANACONDA PROMPT**



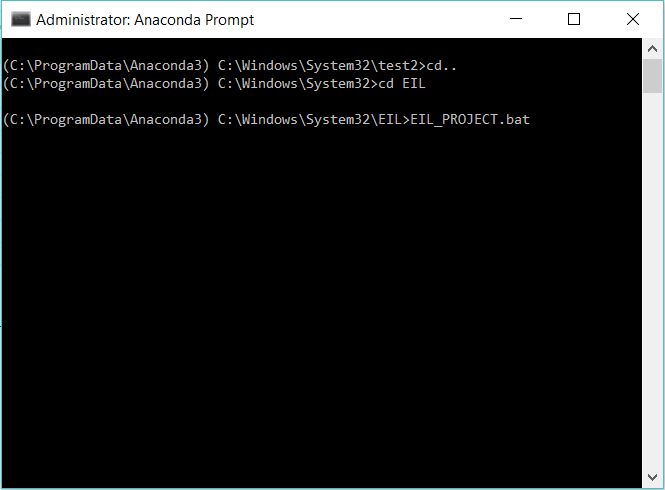
**14.2 OPENING JUPYTER NOTEBOOK**



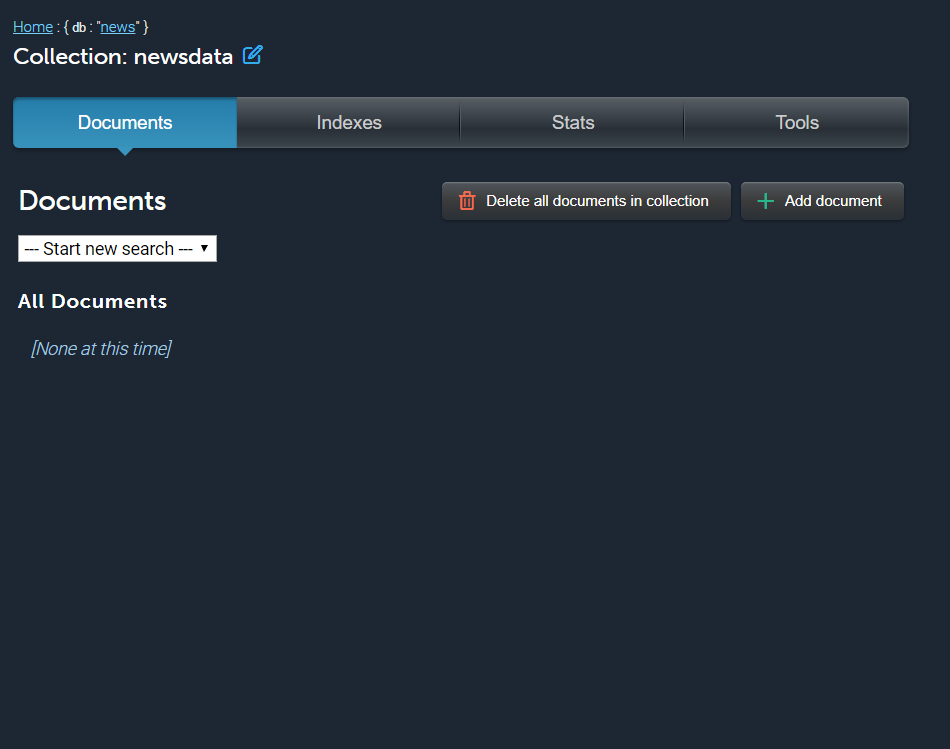
**14.3 PROJECT DIRECTORY**



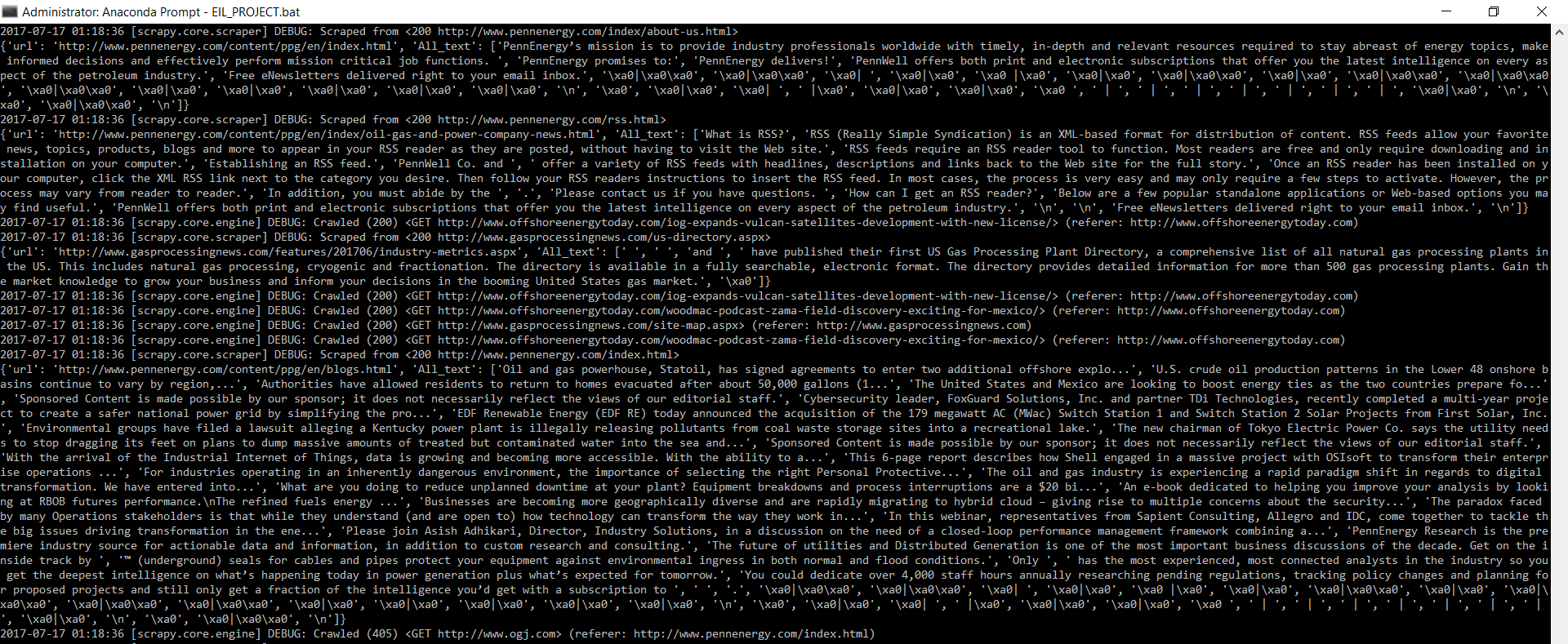
**14.4 RUNNING THE BATCH FILE**



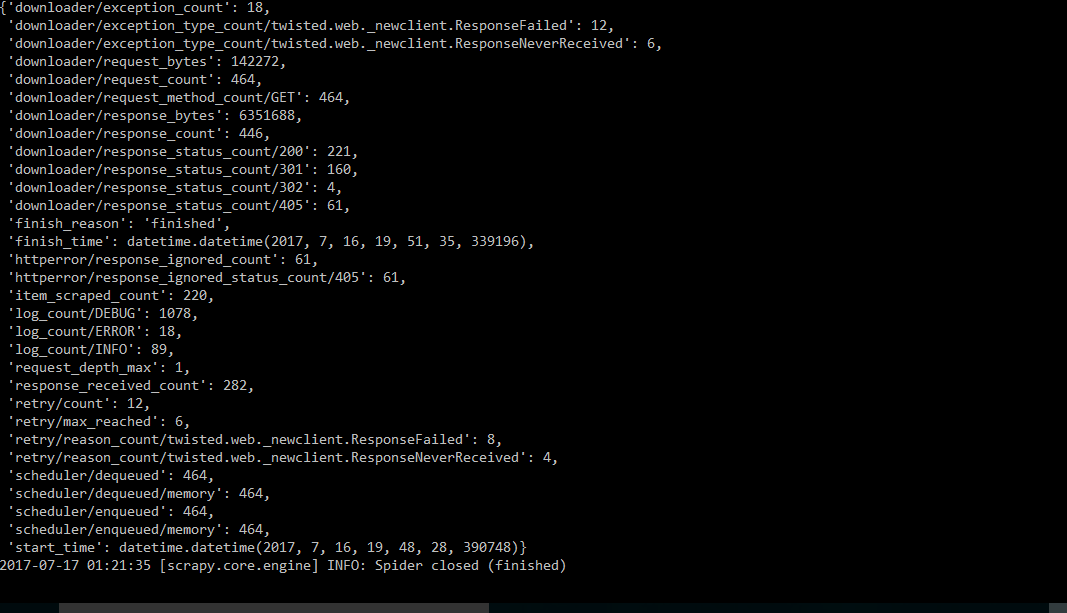
**14.5 mLAB DATABSE BEFORE EXECUTION**

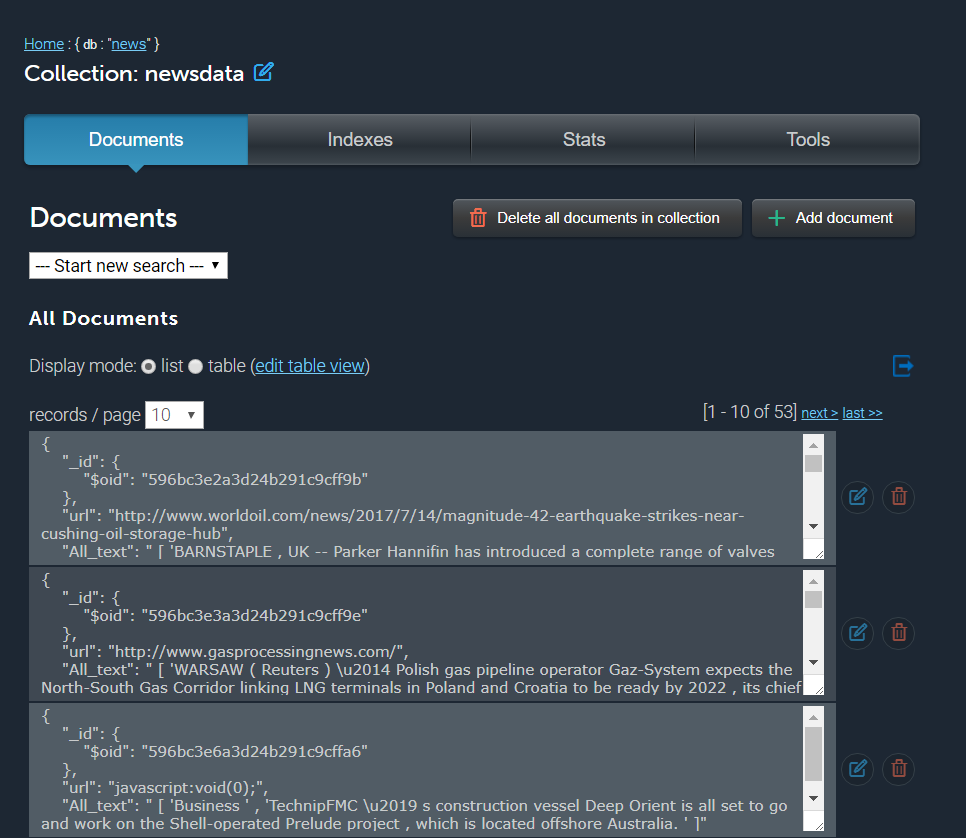


**14.6 CRAWLING WEBSITES**



**14.7 APPLYING NLP AND STORING IT ON ONLINE DATABASE**

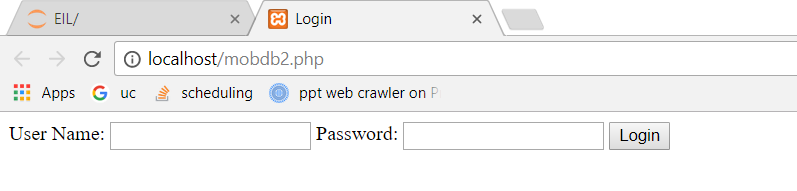


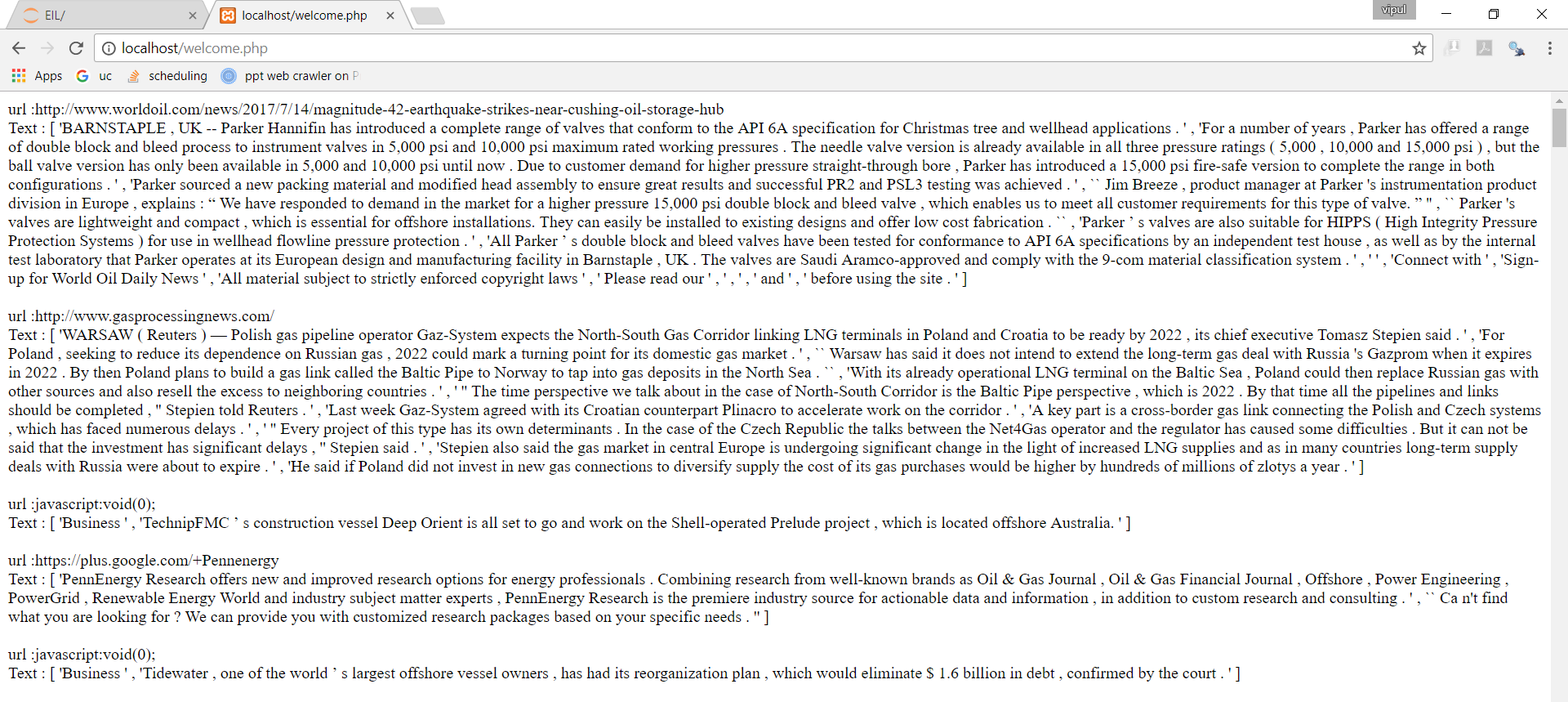


**NOTE:**

**220 data items were scraped from a single website whereas the database only shows 53 entries compromising of text from all the websites. This shows that has been processed before storing.**

**14.8 DISPLAYING DATA**





**14.9 AUTOMATION**



**15.CONCLUSION**

The corporate strategy in terms of various business opportunities in the core areas of operation / strength were analyzed with ETL and made available to the decision makers. The outcome of the analysis is as follows:

1. Business opportunities in the core field
2. Enhancement of existing facilities
3. Information regarding competitors
4. New / planned projects

These analysis is carried out online with a time gap of 1/4 an hour. The leverages the decision makers to make better and effective corporate strategy.

**16 FURTHER SCOPE**

* This project can be transformed into a **MVP**(MINIMUM VIABLE PRODUCT) which crawls and process data according to the user specified keywords and not pre-coded keywords.
* More spiders can be built and joined to the present crawler so as to crawl more websites.
* A proper bot can be developed which can handle robots of different website so that crawling of data is easier.

**17. BIBLIOGRAPHY**

* http://mherman.org/blog/2012/11/05/scraping-web-pages-with-scrapy/#.WW0KZIiGNPY
* https://elitedatascience.com/learn-python-for-data-science
* https://elitedatascience.com/python-web-scraping-libraries
* http://www.nltk.org/
* https://stackoverflow.com/
* https://docs.continuum.io/anaconda-cloud/
* https://www.youtube.com/
* https://en.wikipedia.org/
* https://mlab.com/
* http://docs.mlab.com/
* https://api.mongodb.com/python/current/
* https://www.google.co.in