**Rail Info-graphics and Prediction System**

*Submitted in partial fulfilment of the requirements of*

*the degree of*

**BACHELOR OF TECHNOLOGY**

National Institute of Technology Delhi

*by*

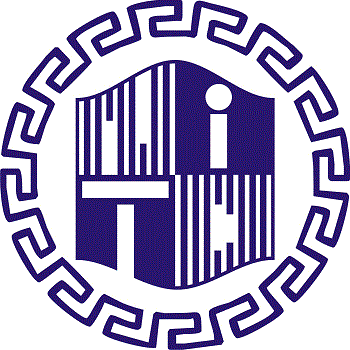
Pedapati Gnanadeep (141100041)

Shruti Bhadoriya (141100007)

*Under the guidance of*

Ms.Pooja Gupta

Assitant Professor, NIT Delhi



**Department of Computer Science Engineering**

NATIONAL INSTITUTE OF TECHNOLOGY DELHI

2017 - 2018

**Acknowledgement**

The satisfaction that is generated by the successful completion of a task would remain unfulfilled without mentioning people who have encouraged and guided us at every step towards the completion of the task. First, we would like to extend our sincere thanks to our project mentor **Ms. Pooja Gupta,** for her guidance and support. Throughout the project she always gave her valuable advice and suggestions for the betterment of our work.

This opportunity gave us a chance to sharpen our working methodology to a higher extent and to solve the problems in a better and easy way, so that it can be presented in a better and understandable manner.

Our heartfelt gratitude also goes to **Dr. Anurag Singh**, Head of the Department, Computer Science Department, National Institute of Technology Delhi, for providing us with the opportunity to avail the excellent facilities and infrastructure of the institute.

Shruti Bhadoriya(141100007)

Pedapati Gnanadeep(141100041)

**Declaration**

I declare that this written submission represents my ideas in my own words and where other’s ideas or words have been included, I have adequately cited and referenced the original sources, I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea /data / fact/ source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not properly cited or from whom proper permission has not been taken when needed.

Name:Pedapati Gnanadeep Ms. Pooja Gupta

Roll no:141100041 Assistant Professor

Date: 30/11/2017

**Declaration**

I declare that this written submission represents my ideas in my own words and where other’s ideas or words have been included, I have adequately cited and referenced the original sources, I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea /data / fact/ source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not properly cited or from whom proper permission has not been taken when needed.

Name:Shruti Bhadoriya Ms. Pooja Gupta

Roll no:141100007 Assistant Professor

Date: 30/11/2017

**List of Figures**

**Page no**

* **Figure 1:** Philosophy of MAQ Software xii
* **Figure 2:** Graphical Illustration of Agile Model 2
* **Figure 3:** Power BI Connection 5
* **Figure 4:** Power BI report 6
* **Figure 5:** Bar Chart 7
* **Figure 6:** Stacked Bar Chart 7
* **Figure 7:** Line Chart 8
* **Figure 8:** Slicer 8
* **Figure 9:** Pie Chart 8
* **Figure 10:** Web Driver Architecture 11
* **Figure 11:**  ETL process 15
* **Figure 12:** Multidimensional Cube 22
* **Figure 13:** Snippet of Schema 22
* **Figure 14:**  Power BI report 2 23

**List of Abbreviation**

* **LOB:** Line of Business
* **BI:** Business Intelligence
* **SDLC:** Software Development Life Cycle
* **SQL:** Structured Query Language
* **T-SQL :** Transact Structured Query Language
* **CTE:** Common Table Expression
* **CSV :**Comma Separated Values
* **API:** Application Program Interface
* **SSIS :** SQL Server Integration Services
* **SSMS:**  SQL Server Management Studio
* **WORA:** Write once, run anywhere
* **ETL:** Extract Transform and Load
* **DOM:** Document Object Model
* **HTML:** Hyper Text Markup Language
* **CSS:** Cascading Style Sheet
* **UI:** User Interface
* **JAR:** Java Archive
* **URL-**Uniform Resource Locator
* **ODS**-Operational Data Store
* **SA**-Staging Area
* **CDW**-Central Data Warehouse

**Chapter 1**

**Introduction**

Indian Railways owing to its cost-efficiency remains the most convenient and sought-after mode of transportation. Indian railways is seventh largest employer in the world according to Forbes, but the question arises, in spite of the voluminous number of employees, does passengers have access to sufficient information to make their travel hassle free? Time and again Indian Railways have been critically questioned on the failure to handle the large run-time delays and the latent factors behind the cause.

* 1. **Motivation**

To cite the official figures in the year 2012-13,on any given day, there were 12617 trains running which carried 8.42 billion i.e. 842.1 crore passengers where one could find that 23.07 million passengers travelled on daily basis. This accounts for 1.9 % of Indian population travelling at any given day of the year. The number of people adopting railways as their mode of commute is massive. Frequent delays and lack of information combined is a principal problem confronted by the passengers.

Other major problems encountered by the passengers is to make a decision about the train they would opt for, based on the delay, seat availability, priority, amount of traffic, connectivity and other additional factors not to be overlooked. We have observed that the data that is present accounts for only a past few days of record and reaching to the conclusion would often require intensive search thus consuming a lot of time.

The delay factor substitutes one of the major factor for selecting a particular train for the desired route, since it is not always possible to find a train to a given destination. In such scenarios, we then try to explore other options by searching the connecting train to the same destination via an intermediate stop. If we are not equipped with sufficient trend in previous delay, there are chances that you might miss the train.

One of the major challenge that we face is the collection of data, the historical data is not available directly at one source, from where it could be obtained. The website and sources currently operating have not yet fulfilled the objective that we aim to implement. We have thoroughly examined various sources and found them to be lacking in terms of volume of data, wieldy and straightforward representation of this data and prediction in the same regard.

Our aim is to collect relevant information, analyse the data and create visual based report which are easier to comprehend and would expedite the decision making process. The interactive visual reports would be of great assistance to the user, but at times the user might not be equipped with sufficient ability to apprehend the visual reports to arrive at conclusion. For such scenarios we aim to simplify the task further by making use of the data at our disposal to predict or suggest the train to the user measured on various metrics like delay, type of train etc.

**1.2 Problem Formulation**

Usually for a given route, there are many trains that travel via the same route as desired by the person, but this necessarily doesn’t mean that all the trains travelling on same route are going to take same time, some of them might get delayed frequently while for some other delay might be seldom. The main goal of the project is to develop application which would be helpful to the targeted audience in selecting better alternative from the available options. We intend to develop interactive visual based reports which can be customized as per the need using Power BI along with prediction of the best trains for a given route using machine learning and artificial intelligence.

**Chapter 2**

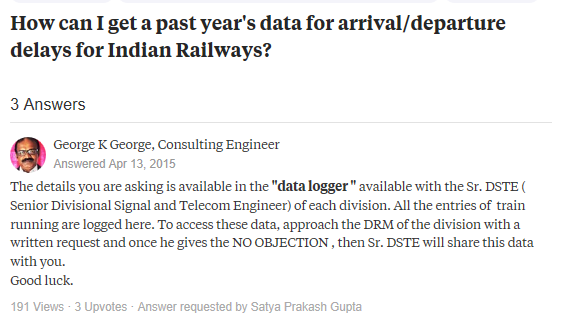
**Literature Review**

Link Set 1: Official figures released by Indian Railways

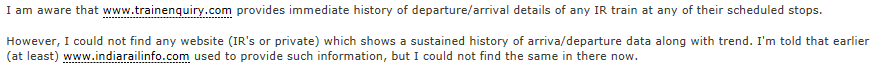
Link1.1:<http://www.indianrailways.gov.in/railwayboard/view_section.jsp?id=0%2C1%2C304%2C366%2C554%2C1451%2C1454&lang=0>

Link Set 2: investigation about the presence of historical data

Link2.1:<https://www.quora.com/How-can-I-get-a-past-years-data-for-arrival-departure-delays-for-Indian-Railways>



Link2.2:<http://www.indiamike.com/india/indian-railways-f10/indian-railway-ontime-performance-train-delay-t139389/>

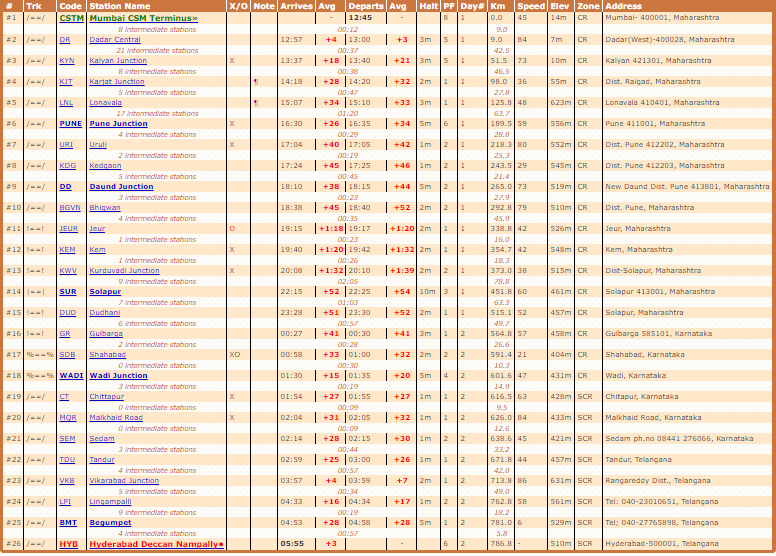


Link2.3: <https://www.quora.com/How-can-I-get-statistical-data-related-to-Indian-railways-of-last-5-or-10-year>s

Link Set 3: Project work with a slight resemblance with the project that we are currently implementing

Link 3.1

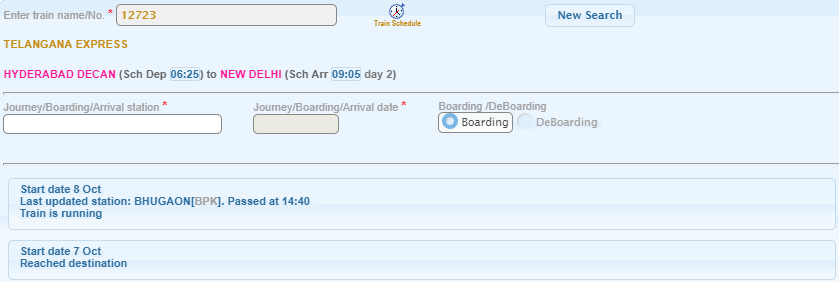
IndianRailInfo: <https://indiarailinfo.com/>



Data displayed on this site might not be precise. Moreover, it shows the data corresponding to average delay only.

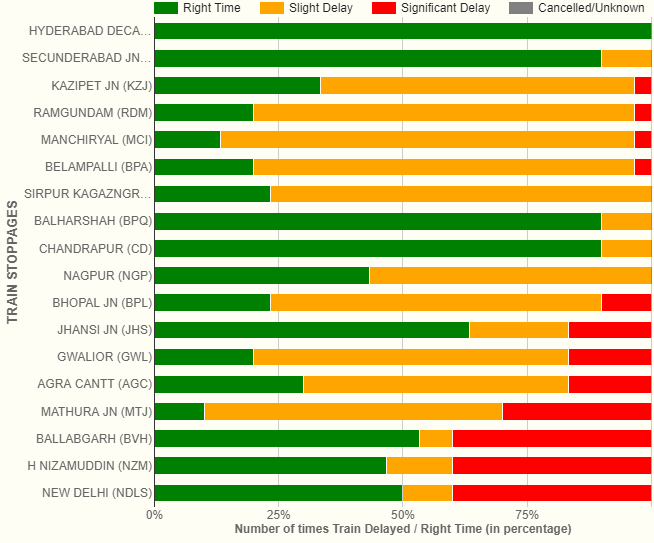
Link 3.2

NTES: <https://enquiry.indianrail.gov.in/ntes/>



Provides the running instances corresponding to past two to three runs only.

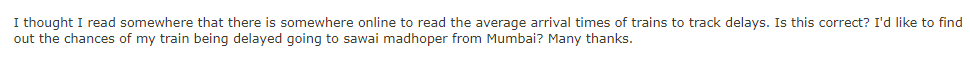
**Link 3.3** : Etrain:[https://etrain.info/in?PAGE=runningHistory--12349--1m#!PAGE=runningHistory--12723--1m](https://etrain.info/in?PAGE=runningHistory--12349--1m" \l "!PAGE=runningHistory--12723--1m)



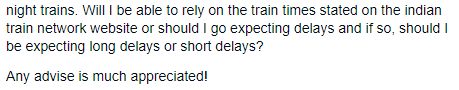
Provides number of times the train has been delayed but with no precise details.

Link Set 4: People enquiring on public forums about the delay trends and delay information

Link 4.1: <http://www.indiamike.com/india/indian-railways-f10/average-train-delays-t144109/>



Link 4.2: <https://www.tripadvisor.in/ShowTopic-g293860-i511-k7331190-Do_Indian_trains_run_on_time-India.html>



Miscellaneous:

<http://dev3.acmdev.org/posters/dev-final15.pdf> -research paper trying to investigate the reason behind the delay caused.

**Chapter 3**

**Tools and Technologies Used**

Project Rail Info-Graphics and Prediction System would be implemented in two parts:

* First Phase is where the Info-Graphics for the relevant data would be implemented making use of business intelligence.
* While the second phase would utilise the data to make prediction using machine learning.

In order to implement the first phase the several tools, technologies and framework had been used. Following is the description of the same.

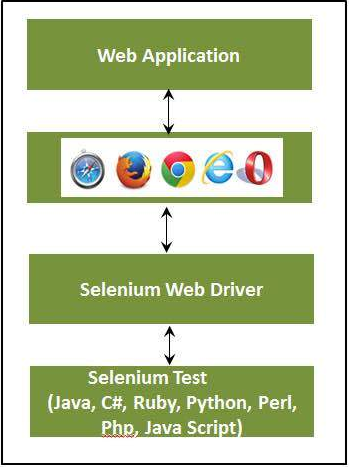
**3.1 Selenium Web Driver**

Selenium web driver was significantly used during the project, we mostly worked on the part of the project where the data pull process was to be automated. **Selenium** is a portable software-testing framework used for web applications.Selenium is used to automatebrowsers. Basically, it is used as tool for automating web applications for testing purposes, but it is not limited to that only. Web-based administration tasks which tends to be tedious and consuming manpower can be automated as well with the help of selenium.

Selenium web driver is mostly used in the cases where:

* You want to create robust, browser-based regression automation suites and tests.
* You want to scale and distribute scripts across many environments.

 Selenium WebDriver is originally a collection of language specific-bindings which is used to drive browser - the way it is meant to be driven. Selenium Remote Control which has been officially deprecated is the predecessor of Selenium Web Driver.



**Figure 10: Web Driver Architecture**

**3.1.1 Selenium Locators**

Locator is a command that tells Selenium Web Driver which GUI elements (say Text Box, Buttons, Check Boxes etc) its needs to operate on. Different types of locators have been mentioned below.

**3.1.1.1 Locating by ID**

As the ID’s are unique for each element ,so this is the most common method to locate elements.

**Format:**id=elements’ id

**3.1.1.2 Locating by Name**

Using this method to locate the element is almost same as that of locating by ID except for, the prefix “name=” is used instead of “id=”.

**Format:**name=element’s name

**3.1.1.3 Locating by Link Text**

This is used in case where hyperlinks are present. In order to access the link text we select the string between anchor tags which is the hyperlink text and is prefixed by “link=”.

**Format**: link=link\_text

**3.1.1.4 Locating by CSS Selector**

**CSS Selectors are string patterns which are useful in identification of elements using different combinations of HTML tag like id, attribute, class and name.**

Commonly used combination of css selector is discussed below.

* Tag and ID
* Tag and class
* Tag and attribute
* Tag, class, and attribute
* Inner text

**3.1.1.5 Locating by DOM (Document Object Model)**

The Document Object Model (DOM), refers as to how the HTML webpage and its elements are strucured. Selenium can use the DOM to access page elements.

There are four ways to access an element via DOM:

* getElementById
* getElementsByName
* dom:name
* dom:index

**3.1.1.6 Locating by Xpath**

XPath is the language used when locating XML (Extensible Markup Language) nodes. Since HTML can be thought of as an implementation of XML, we can also use XPath in locating HTML elements. It can access almost any element, even those without class, name, or id attributes.

There are two types of xpath

* Absolute XPath
* Relative Xpath

**3.1.2 Implicit wait and Explicit wait in Selenium**

**3.1.2.1 Why wait is required in Selenium?**

Most of the web applications are developed using Ajax and Javascript. Different elements in a page may load at different time ,at the time when webpage is loaded.

This makes it difficult to locate elements as each element has different loading time and in case if the element is not found then the exception will be thrown “ElementNotVisibleException”. This problem is solved by the use of wait in the program which would wait as per the conditions defined by the programmer.

**3.1.2.2 Implicit Wait**

In implicit wait the web driver waits for a certain period of time before throwing “NoSuchElemenExcpetion” or “ElementNotVisibleException”.By default the waiting time for the driver is set to 0. Once the time to wait is set then the web driver wait for that amount of time before throwing any exception. Implicit wait condition holds true for the entire time for which the browser is open. Therefore any search for elements on the page would wait for the time for which implicit wait has been imposed.

**3.1.2.3 Explicit Wait**

Explicit wait is defined as the wait time for which the web driver waits for a certain condition to happen before going further in the code. One of the type of explicit wait is where you define for how much time will the browser will wait irrespective of the any condition, it is Thread.sleep(), this can make program slow. There are some methods predefined which help in writing code that would wait only for as long as requirement. WebDriverWait along combination of ExpectedCondition is one of the way in which the task can be accomplished.

Some of the expected condition are listed as follows:

* elementToBeClickable(By locator) : waits until an element is visible and enabled
* elementToBeSelected(WebElement element): waits until an element is selected
* presenceOfElementLocated(By locator) : waits until presence of an element
* textToBePresentInElement(By locator, String text): waits until specific text is present in the an element.

**3.1.3 Why use selenium over other conventional scraping tools?**

* Selenium web driver is capable of pulling data from web pages with dynamic content with much ease.
* With the help of selenium you could use it to send the value to a website i.e. fill a field or an entire form through your program, unlike other tools which are incapable of doing this.
* Most of the conventional data scraping tool operate on Internet explorer, while with selenium you can overcome this limitation, since it is operable on a variety of browsers like internet explorer, google chrome, mozzila firefox, opera etc.

**3.2 Transact-SQL**

Transact SQL or better known as T-SQL is a set of programming extensions from Sybase and Microsoft that add several features to the Structured Query Language (SQL), including transaction control, exception and error handling, row processing and declared variables.

T-SQL is used to interact with relational databases. T-SQL expands on the SQL standard to include procedural programming, local variables, various support functions for string processing, date processing, mathematics, etc. and changes to the DELETE and UPDATE statements.

Transact-SQL is central to using Microsoft SQL Server. All applications that communicate with an instance of SQL Server do so by sending Transact-SQL statements to the server, regardless of the user interface of the application.

T-SQL's transaction and journaling system, handles just about anything - including a power cycle or hardware failure - without database corruption, and if something gets messed up it fixes it automatically.

T-SQL support CTE. A common table expression (CTE) can be thought of as a temporary result set that is defined within the execution scope of a single SELECT, INSERT, UPDATE, DELETE, or CREATE VIEW statement. A CTE is similar to a derived table in that it is not stored as an object and lasts only for the duration of the query. Unlike a derived table, a CTE can be self-referencing and can be referenced multiple times in the same query. It simplifies complex queries and most importantly enables you to use recursion.

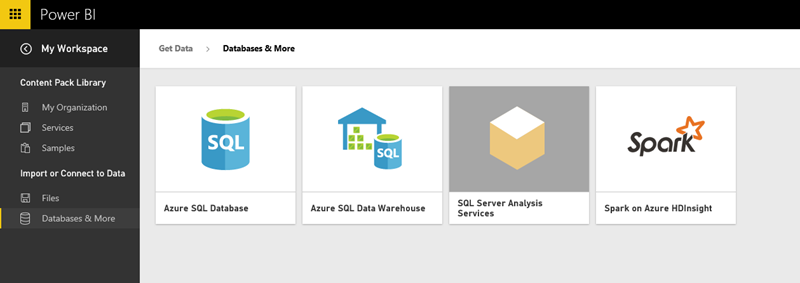
T-SQL is the SQL dialect that the product SQL Server is using. Transact-SQL is central to using SQL Server. All applications that communicate with an instance of SQL Server do so by sending Transact-SQL statements to the server, regardless of the user interface of the application. SQL Server is tied to Transact-SQL (T-SQL), an implementation of SQL from Microsoft that adds a set of proprietary programming extensions to the standard language.

**3.2.1** **Difference between T-SQL and SQL**

* T-SQL adds a number of features that are not available in SQL. This includes procedural programming elements and a local variable to provide more flexible control of how the application flows.
* A number of functions were also added to T-SQL to make it more powerful; functions for mathematical operations, string operations, date and time processing, and the like.
* These additions make T-SQL comply with the Turing completeness test, a test that determines the universality of a computing language. SQL is not Turing complete and is very limited in the scope of what it can do.
* Another significant difference between T-SQL and SQL is the changes done to the DELETE and UPDATE commands that are already available in SQL. With T-SQL, the DELETE and UPDATE commands both allow the inclusion of a FROM clause which allows the use of JOINs. This simplifies the filtering of records to easily pick out the entries that match a certain criteria unlike with SQL.
* SQL is non-procedural language since it deals with what data to be extracted. Whereas T-SQL is procedure language since it deals with what data to be executed and how it should be displayed.
* The SQL queries in SQL are submitted individually to the database server, while in T-SQL the batch program is written where in all commands are submitted to the server in a single go.

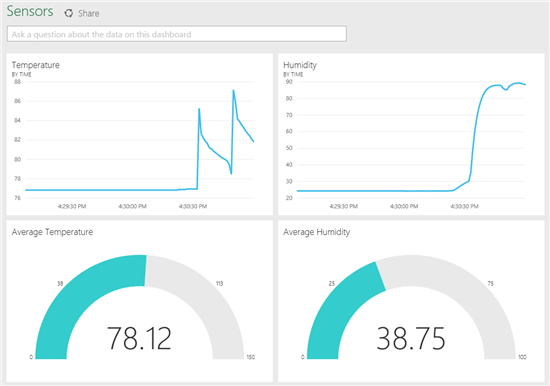
**3.3 Power BI**

Power BI is a cloud-based business analytics service from Microsoft that empowers anyone to experience any data – structured or unstructured – via simple drag-and-drop ease. Unlike many other dashboard solutions, Power BI can render live dashboards with moving charts and continuously updated visualizations for monitoring real-time streams from supported data sources.

****

**Figure 3: Power BI Connection**

As shown in figure 3, we can use the Power BI REST API with any data source or Azure Streaming Analytics to render live Power BI Dashboards automatically. Alternatively, you can get near real-time analytics using simple “direct connect” data sources such as Analysis Services, Azure SQL Database, Azure SQL Data Warehouse or Spark with Power BI Reports.

****

**Figure 4: Power BI report**

As shown in fig 3, We can get any type of visuals from Power BI ,Microsoft PowerBI tools makes the reporting part so easy that we can create our reports in short period of time with so much attractive and correct and real time data, milliseconds refresh data we can retrieve with our job scheduling.

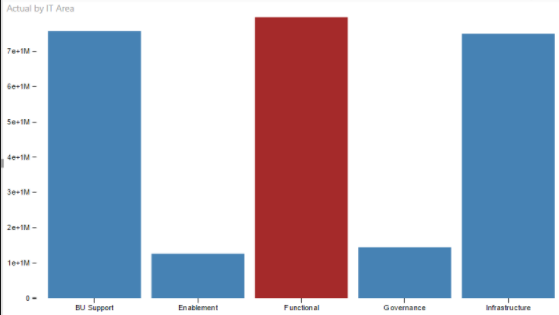
You can import the data which is required to make the power BI report from various sources. It can be an excel sheet, CSV (comma separated value) file, database on your local machine, data in cloud etc.

Power BI has Q&A feature to explore your data using intuitive, natural language capabilities and receive answers in the form of charts and graphs. Q&A is different from a search engine -- Q&A only provides results about the data in Power BI. Data visualizations (aka visuals) helps us to interact with data to find business insights.

**3.3.1 Power BI visualizations**

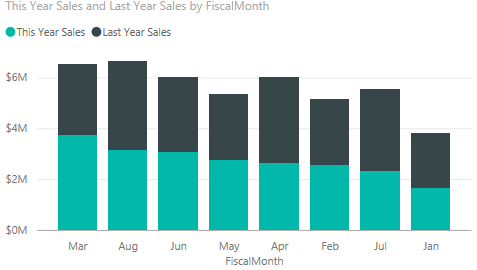
The types of visualization generally used in Power BI reports are:

1. Bar Chart



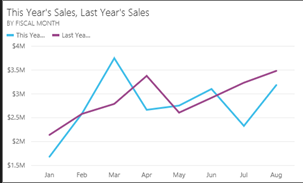
**Figure 5: Bar chart**

1. Stacked Bar Chart



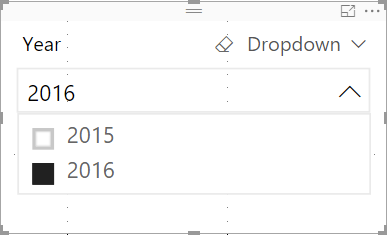
**Figure 6: Stacked Bar Chart**

1. Line Chart



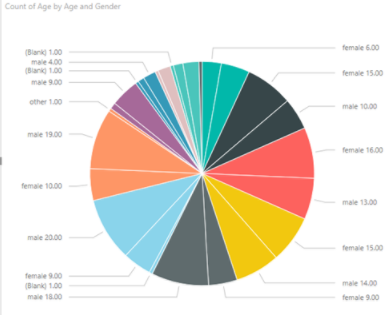
**Figure 7:Line Chart**

1. Slicer



**Figure 8: Slicer**

1. Pie Chart



**Figure 9: Pie Chart**

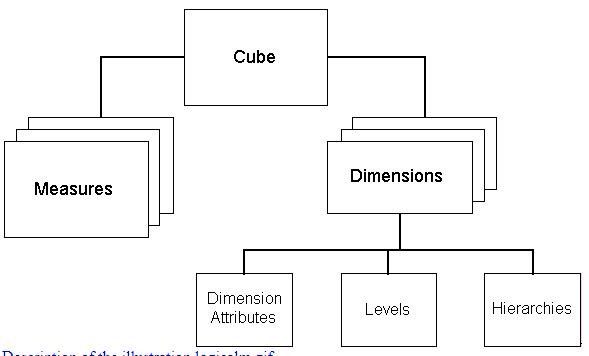
**3.4 Data Cubes**

Data cubes are multi-dimensional database, in other word they are extension to the 2 dimensional relational table. Cubes can be 3 dimensional, 4 dimensional etc depending on the requirement.

OLAP (OnLine Analytical Processing)make use of cubes since they provide deeper insights to the data. The data is present in dimension tables and the fact table then summarizes the data in dimension table for a given attribute, so we can say that fact table stores the data in aggregated form which are often called measures. MDX (multi-dimensional expressions) queries are written to extract data from cubes.

Since the data is stored in aggregated form, this is often useful for the analyst to perform data analysis, establish trends, measure performance. We don’t perform calculation for the data in cubes , since the data in the cube has already once been analysed, processed and aggregated into the form of the cube. This implies that the data in the cube is historical and not dynamic and real-time data. These characteristics of a data cube perfectly aligns with the reporting purposes where millions of records are to be processed at a time. A data cube is a single entity where all the data for analytics purpose could be found without having to refer to different relational tables and their relations.

**3.4.1 Multidimensional Model**



As can be seen from the figure above, multidimensional model consists of dimensions and measures (which are present in fact table).

Dimensions has attribute that are unique for a feature that categorizes the data. While in fact tables, only the integer aggregated values i.e. measures and foreign keys are present. Fact table contain data that is relevant in decision making process.

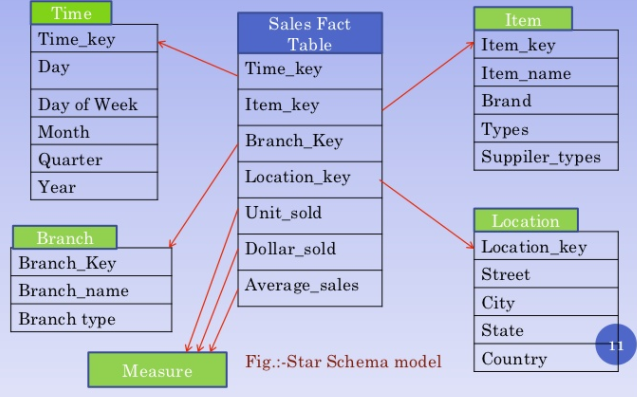
**3.4.2 Schema**

There are different types of implementation of the data cubes:

* Star Schema Model
* Snow Flake Schema
* Fact Constellation

**3.4.2.1 Star Schema**

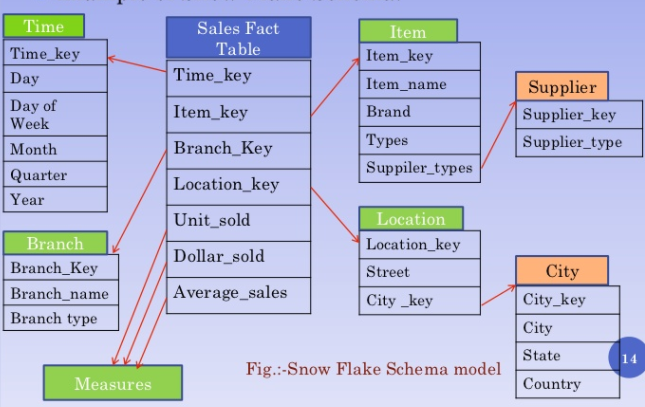
It is called star schema because this schema is similar to a star where all the relationship originates from a single point source. It is similar to inner join between a fact table and multiple dimension tables. Dimension has a primary key while the fact has a foreign key. The point to be noted in star schema is that the dimension table are not related to one another, they all are only related to fact table. Performance in this schema is highly optimized.



**3.4.2.2 Snowflake Schema**

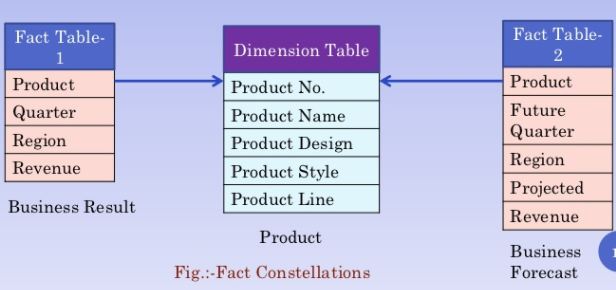
This is similar to star schema except for the fact that in this the dimension table are normalized so that the hierarchy is created and to get rid of redundancy and anomalies.

As there is no redundancy in data, the denormalized schema then occupy less space on disk. Due to normalization, the queries to extract data from snow flake model becomes quite complex when compared with queries in star-schema model.



**3.4.2.3 Fact Constellations**

As the name suggests, it is a set of fact tables which have shared dimensions among themselves. Multiple star schema can share a fact constellation. This schema is more complex as compared to the other two schema, since it has more than one fact table with shared data which makes it hard to manage data and data relationship. Several aggregation leads to enhanced complexity.



**3.5 Java**

**Java** is high level programming which is concurrent, object-oriented, and has been designed specifically to have minimum possible implementation dependencies. It is "write once, run anywhere" (WORA), application that is being used by developers which means the  compiled Java code can run on all platforms which support Java without the need for recompiling the code everytime.

**3.6 SSIS (SQL Server Integration Services)**

**SQL Server Integration Services** (**SSIS**) is a component of the Microsoft SQL Server database software that can be used to perform a broad range of data migration tasks. SSIS is a platform for data integration and workflow applications. It has a data warehousing tool which is utilized for data extraction, transformation, and loading commonly termed as ETL. This tool might be used in order to automate maintenance of SQL Server databases and updating them to multidimensional data cubes.

**3.7 SSMS(SQL Server Management Studio)**

**SQL Server Management Studio** (SSMS) is a software application first launched with Microsoft SQL Server 2005 that is used for configuring, managing, and administering all components within Microsoft SQL Server. This tool has both script editors and graphical tools which work with objects and features of the server.