Project On

Airline Analysis

A Report for Project 2

Bachelor of Technology in Computer Science and Engineering

Under the Guidance of

Name of the Guide

Mr. Anuj Kumar Bharti

Submitted by

Name of the Student :

Shubham Jha

16SCSE101872

1613101719

Department of Computer Science and Engineering

GALGOTIAS UNIVERSITY GREATER NOIDA UP.

Table of contents

Abstract

In the contemporary world, Data analysis is a challenge in the era of varied inters- disciplines though there is a specialization in the respective disciplines.

In other words, effective data analytics helps in analyzing the data of any business system. But it is the big data which helps and axial rates the process of analysis of data paving way for a success of any business intelligence system. With the expansion of the industry, the data of the industry also expands. Then, it is increasingly difficult to handle huge amount of data that gets generated no matter what's the business is like, range of fields from social media to finance, flight data, environment and health.

Big Data can be used to assess risk in the insurance industry and to track reactions to products in real time. Big Data is also used to monitor things as diverse as wave movements, flight data, traffic data, financial transactions, health and crime. The challenge of Big Data is how to use it to create something that is value to the user.

How can it be gathered, stored, processed and analyzed it to turn the raw data information to support decision making. In this paper Big Data is depicted in a form of case study for Airline data.

The proposed method is made by considering following scenario under consideration

An Airport has huge amount of data related to number of flights, data and time of arrival and dispatch, flight routes, No. of airports operating in each country, list of active airlines in each country. The problem they faced till now it's, they have ability to analyze limited data from databases. The Proposed model intension is to develop a model for the airline data to provide platform for new analytics based on the following queries.

Introduction

1.1 Overall description

Big Data is not only a Broad term but also a latest approach to analyze a complex and huge amount of data; there is no single accepted definition for Big Data. But many researchers working on Big Data have defined Big Data in different ways. One such approach is that it is characterized by the widely used 4 V's approach. The first "V" is Volume, from which the Big Data comes from. This is the data which is difficult to handle in conventional data analytics. For example, Volume of data created by the BESCOM (Bangalore Electricity Supply Company) in the process of the power supply and its consumption for Bangalore city or for the entire Karnataka State generates a huge volume of data. To analyze such data, it is the Big data that comes to aid of data analytics; the second "V" is velocity, the high speed at which the data is created, processed and analyzed; the third "V" is variety which helps to analyze the data like face book data which contains all types of variety, like text messages, attachments, images, photos and so on; the forth "V" is Veracity, that is cleanliness and accuracy of the data with the available huge amount of data which is being used for processing.

Researchers working in the structured data face many challenges in analyzing the data. For instance the data created through social media, in blogs, in Facebook posts or Snap chat. These types of data have different structures and formats and are more difficult to store in a traditional business data base. The data in big data comes in all shapes and formats including structured. Working with big data means handling a variety of data formats and structures. Big data can be a data created from sensors which track the movement of objects or changes in the environment such as temperature fluctuations or astronomy data. In the world of the internet

of things, where devices are connected and these wearable create huge volume of data. Thus big data approaches are used to manage and analyze this kind of data. Big Data include data from a whole range of fields such as flight data, population data, financial and health data such data brings as to another V, value which has been proposed by a number of researcher i.e., Veracity.

Most of the time social media is analyzed by advertisers and used to promote produces and events but big data has many other uses. It can also been used to assess risk in the insurance industry and to track reaction to products in real time. Big Data is also used to monitor things as diverse as wave movements, flight data, traffic data, financial transactions, health and crime. The challenge of Big Data is how to use it to create something that is value to the user. How to gather it, store it, process it and analyze it to turn the raw data information to support decision making.

Hadoop allows to store and process Big Data in a distributed environment across group of computers using simple programming models. It is intended to scale up starting with solitary machines and will be scaled to many machines. But now since huge amount of data in Terabytes which is injected into Hadoop Distributed File System files and processed by HDFS Tool.

An Airport has huge amount of data related to number of flights, data and time of arrival and dispatch, flight routes, No. of airports operating in each country, list of active airlines in each country. The problem they faced till now it's, they have ability to analyze limited data from databases. The Proposed model intension is to develop a model for the airline data to provide platform for new analytics based on the following queries.

1.1 Problem Statement

- ✓ Big amount of data generated on hourly basis.
- ✓ A single twin engine aircraft with an average 12 hour flight time can produce up to 844 TB of data
- ✓ There are many active users of flights
- ✓ Many flights are scheduled everyday

- ✓ User varies from common man to celebrities

The proposed method is made by considering following scenario under consideration .An Airport has huge amount of data related to number of flights, data and time of arrival and dispatch, flight routes, No. of airports operating in each country, list of active airlines in each country. The problem they faced till now it's, they have ability to analyze limited data from databases. The Proposed model intension is to develop a model for the airline data to provide platform for new analytics based on the following queries.

1. Extract unstructured data using python language.

2. Make unstructured data into structured using hadoop.

3. Analyse data for the following queries

a) List of airports operating in the country India?

b) How many active airlines in United State.?

c) List of airlines operating with code share?

d) Which country having highest Airport?

e) How many flight having same air code for flight which uses code share?

1.2 Purpose

The main purpose of the project to explore detailed analysis on airline data sets such as listing airports operating in the India, list of airlines having zero stops, list of airlines operating with code share which country has highest airports and list of active airlines in united states. The main objective of project is the processing the big data sets using map reduce component of hadoop ecosystem in distributed environment.

1.3 Motivation and scope

Airline data analysis can provide a solution for businesses to collect and optimize large datasets, improve performance, improve their competitive advantage, and make faster and better decisions.

- ✓ By using airline data analysis, we can save time of users.
- ✓ The data could even be structured, semi-structured or unstructured.
- ✓ Cost savings
- ✓ Implementing new strategies
- ✓ Fraud can be detected the moment it happens

1.4 Operating Environment or Software Environment

Software environment is the term commonly used to refer to support an application. A software environment for a particular application could include the operating system, the database system, specific analysis tools.

The software and hardware that we are using in our project Airline data analysis are:

1.4.1 Intel core i3 and above

1.4.2 Windows 10

1.4.3 Windows subsystem for Linux

1.4.4 Ubuntu

1.4.5 Java JDK 1.8

1.4.6 Hadoop 3.0.0

1.4.7 Map reduce

1.4.8 Microsoft Excel

1.4.9 Minimum RAM 4GB and abov

1.5 Assumptions and Dependencies

Constraints are limitations which are outside the control of the project. The Project must be managed within these constraints.

Assumptions are made about events, or facts outside the control of project. External dependencies are activities which need to be completed before an internal activity can proceed.

Constraints, assumptions and dependencies can create risks that the project may be delayed because access is not provided to the site (assumption).

Assumption will be that the complexity may arise due to large unstructured data set.

1.6 Constraints

Hardware limitation and timing constraints.

High feature may not correspond to semantic similarity.

System Environment

Windows subsystem for Linux with Ubuntu operating system will be required to run the application

Proposed Model

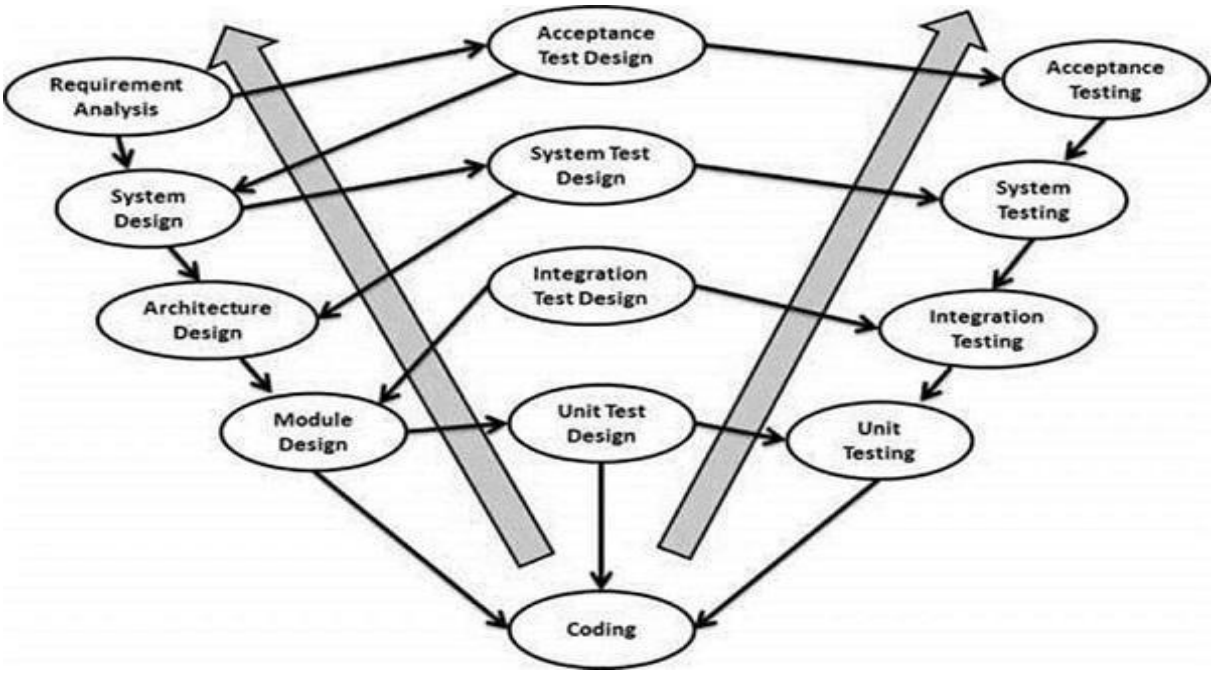
This Project is based on V-model **SDLC** (Software Development Life Cycle)

The V-model is an SDLC model where execution of processes happens in a sequential manner in a V-shape. It is also known as Verification and Validation model.

The V-Model is an extension of the waterfall model and is based on the association of a testing phase for each corresponding development stage. This means that for every single phase in the development cycle, there is a directly associated testing phase. This is a highly-disciplined model and the next phase starts only after completion of the previous phase.

Under the V-Model, the corresponding testing phase of the development phase is planned in parallel. So, there are Verification phases on one side of the 'V' and Validation phases on the other side. The Coding Phase joins the two sides of the V-Model.

The following illustration depicts the different phases in a V-Model of the SDLC.


Requirement Analysis
Acceptance Test Design
Acceptance Testing
System Design
System Test Design
System Testing
Architecture Design
Integration Test Design
Integration Testing
Module Design
Unit Test Design
Unit Testing
Coding

References

[1] http://cra.org/ccc/wpcontent/uploads/sites/2/2015/05 /bigdatawhitepaper.pdf

[2] www.ijcsmc.com/docs/papers/June2017/V6I6201764.pdf

[3] https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html

[4] https://www.uml-diagrams.org/index-examples.html.

[5] https://www.researchgate.net/figure/The-MapReduce-architecture-MapReduce-

[6] https://flume.apache.org/

[7] https://www.digitalocean.com/community/tutorials/how-to-install-

hadoop-in-stand- alone-mode-on-ubuntu-16-04

[8] https://www.tutorialspoint.com/sdlc/sdlc_v_model.htm

[9] https://www.ten10.com/types-testing-introduction-different-types-software-testing