

**Project Report**

**Data Analysis of Flights Traffic**

**using Big Data Hadoop**

**Under Guidance of**

**Prof. R.K. Singh (SCSE)**

**Project By**

Kunwarjeet Singh 12MSE1086

**Abstract**

In this project we will be doing analysis on Big Datasets of Us Flight Traffic using Hadoop Technologies such as MapReduce, Hive , Pig ,Scoop, Hbase. First we will learn Hadoop technologies , then we will be finding trendz in dataset by reducing large data into small based on queries. We will finally make graphs , charts etc to show graphical representation of our analysis. This will be helpful for Airports how to manage and reduce flight delays etc.

**Description about Dataset**

Our Dataset contains info on each single flight details of USA from 2002 – 2008. It contains info such as **travel date, arrivaltime, departuretime, delays in arrival in departure time ,tail num ,flight num , manufacturer details, cancelation details** etc.

Size of Dataset is Around **1GB**

**Software Hardware Requirements**

* Hadoop Distributed System (HDFS)
* Linux / Unix like environment
* MySQL
* System with atleast 6 GB RAM

**Hadoop technologies to be used**

* **MapReduce** is a programming model for processing and generating large data sets with a parallel, distributed algorithm on a cluster
* **Apache Hive** is a data warehouse infrastructure built on top of Hadoop for providing data summarization, query, and analysis.
* **Apache Pig** is a platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs, coupled with infrastructure for evaluating these programs.
* **HBase** is an [open source](http://en.wikipedia.org/wiki/Open_source), [non-relational](http://en.wikipedia.org/wiki/Non-relational_database), [distributed database](http://en.wikipedia.org/wiki/Distributed_database) modeled after [Google's](http://en.wikipedia.org/wiki/Google) [BigTable](http://en.wikipedia.org/wiki/BigTable) and written in [Java](http://en.wikipedia.org/wiki/Java_(programming_language))CloudEra.
* Apache **Sqoop** is a tool designed for efficiently transferring bulk data between Apache Hadoop and structured datastores such as relational databases.
* **Tableau** Public is free software that can allow anyone to connect to a spreadsheet or file and create interactive data visualizations for the web

**Procedure Followed**

* Downloaded Data Set in zip format , extract it you get .csv file
* Copy each DataSet to Virtual Machine one by one.
* Put file into Hadoop via ( Hadoop fs –put filename newname )
* Use Hive query to create Table and load data into it

create table 2003file(Year int,Month int,DayofMonth int,DayOfWeek int,DepTime int,CRSDepTime int,ArrTime int,CRSArrTime int,UniqueCarrier String,FlightNum int,TailNum String,ActualElapsedTime int,CRSElapsedTime int,AirTime int,ArrDelay int,DepDelay int,Origin String,Dest String,Distance int,TaxiIn int,TaxiOut int,Cancelled int,CancellationCode int,Diverted int,CarrierDelay int,WeatherDelay int,NASDelay int,SecurityDelay int,LateAircraftDelay int)

Load data inpath 2003.cs into 2003file

* Use Pig to filter , group this structured data according to our need

A= LOAD '2005' using PigStorage(',') AS(Year:int,Month:int,DayofMonth:int,DayOfWeek:int,DepTime:int,CRSDepTime:int,ArrTime:int,CRSArrTime:int,UniqueCarrier:chararray,FlightNum:int,TailNum:chararray,ActualElapsedTime:int,CRSElapsedTime:int,AirTime:int,ArrDelay:int,DepDelay:int,Origin:chararray,Dest:chararray,Distance:int,TaxiIn:int,TaxiOut:int,Cancelled:int);

C= FILTER A by Origin=='MSY' OR Origin=='BHM' OR Origin=='MIA' OR Origin=='MCO' OR Origin=='FLL' OR Origin=='TPA' OR Origin=='PBI' OR Origin=='JAX';

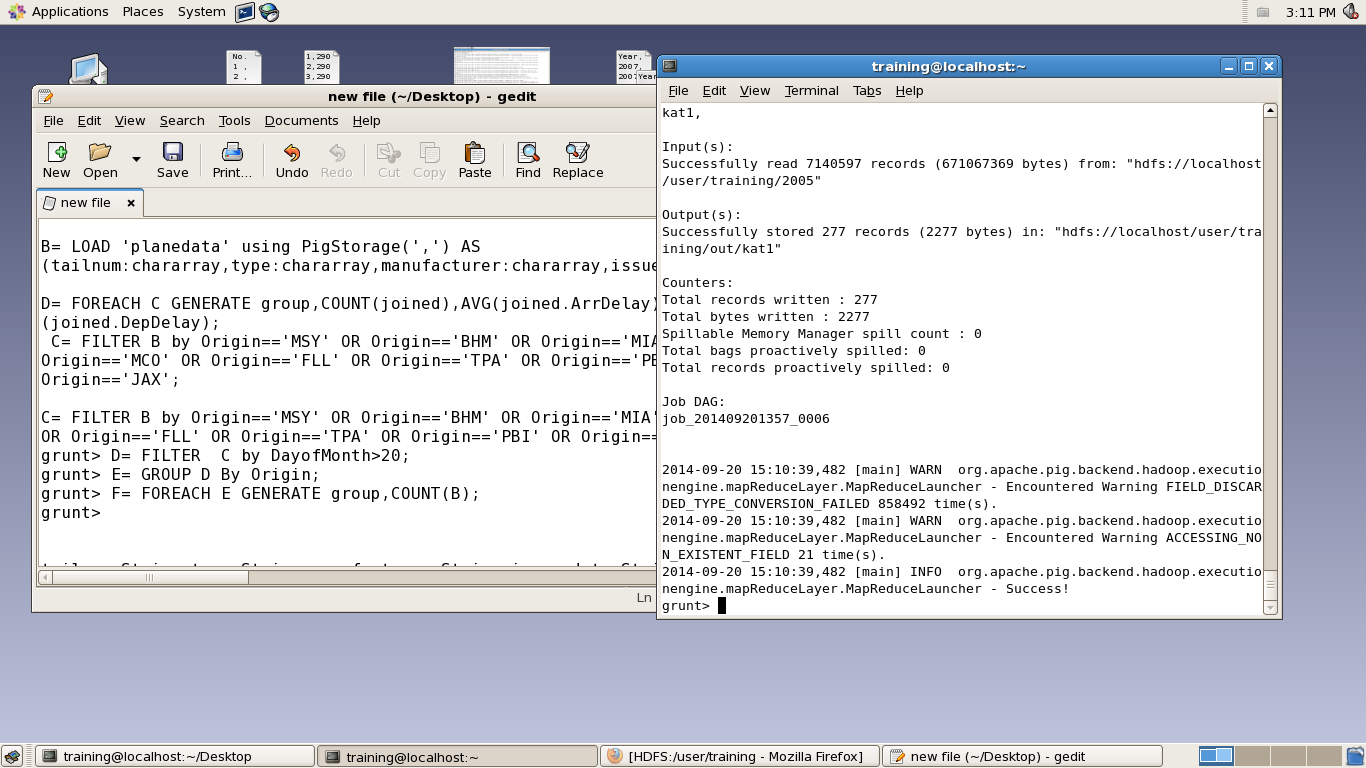
grunt> D= FILTER C by DayofMonth>23

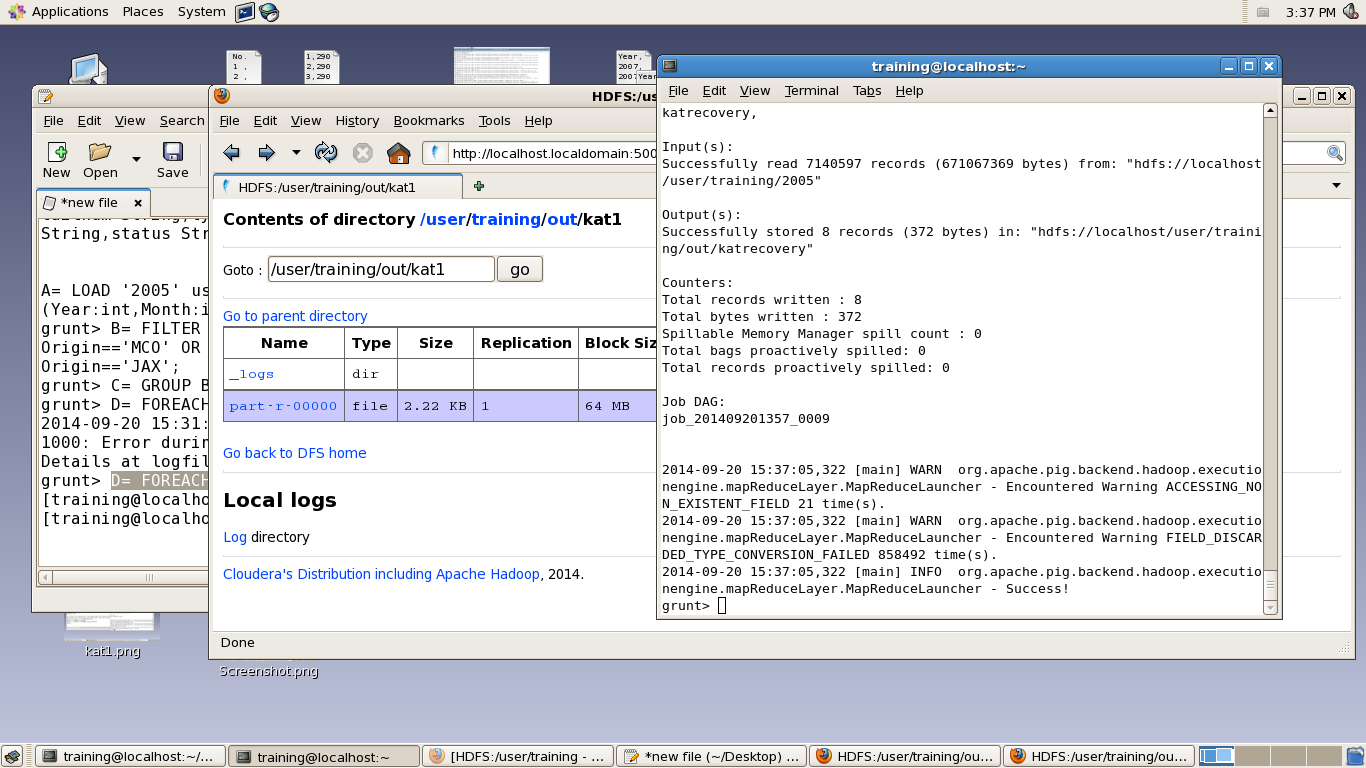
grunt> E= GROUP D By Origin;

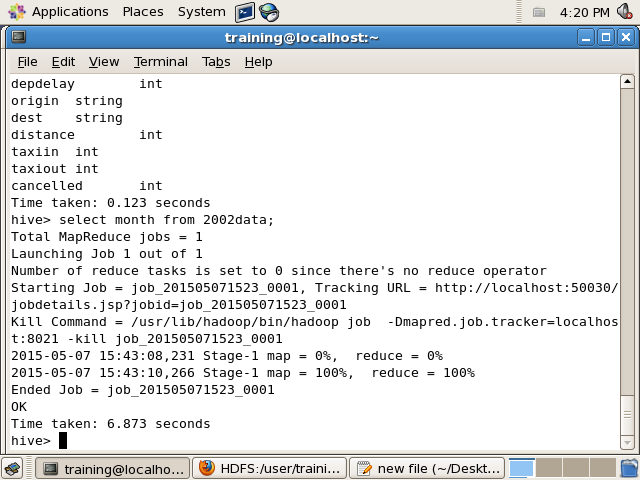
grunt> F= FOREACH E GENERATE group,COUNT(D);

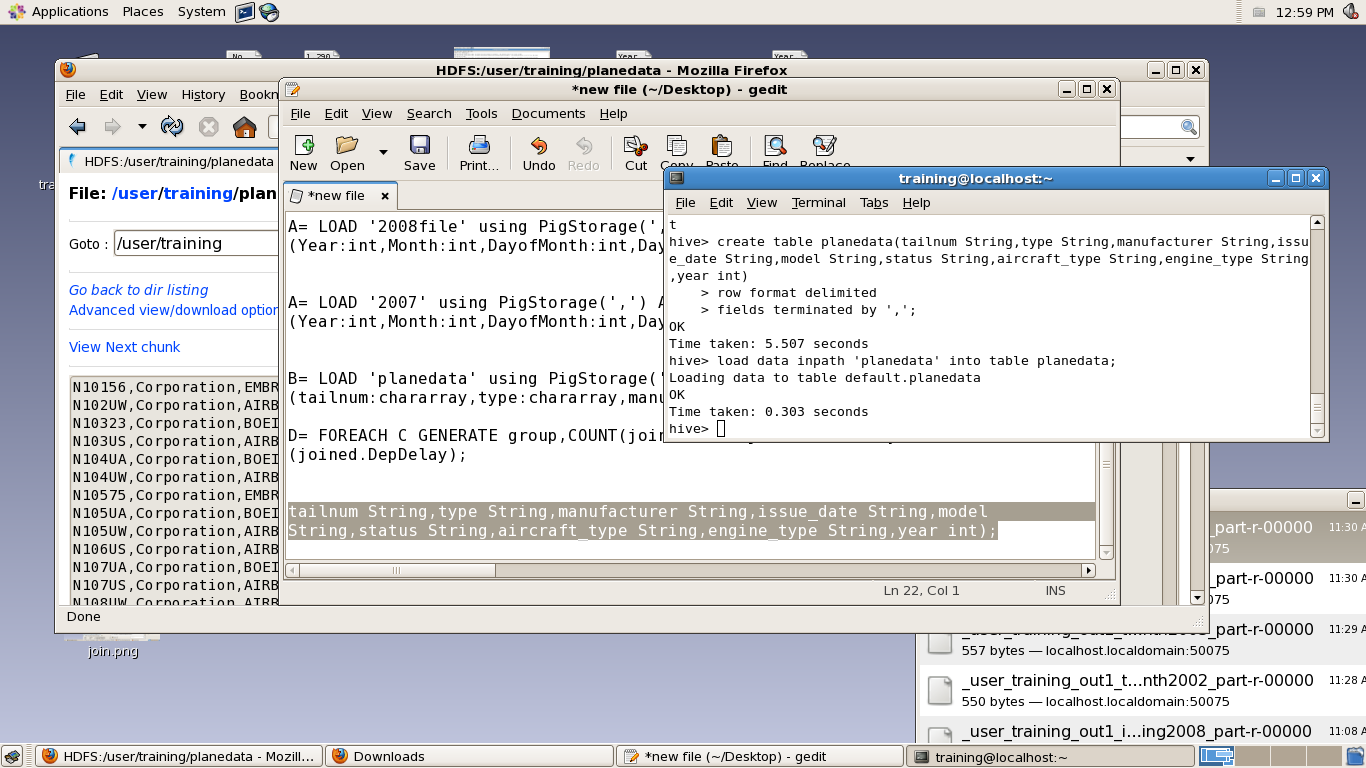
* Dump output - grunt> STORE D into 'out/kat2005';
* Copy output to MS Excel
* Import excel workbook into Tableu Software
* Draw graphs , charts as per need.

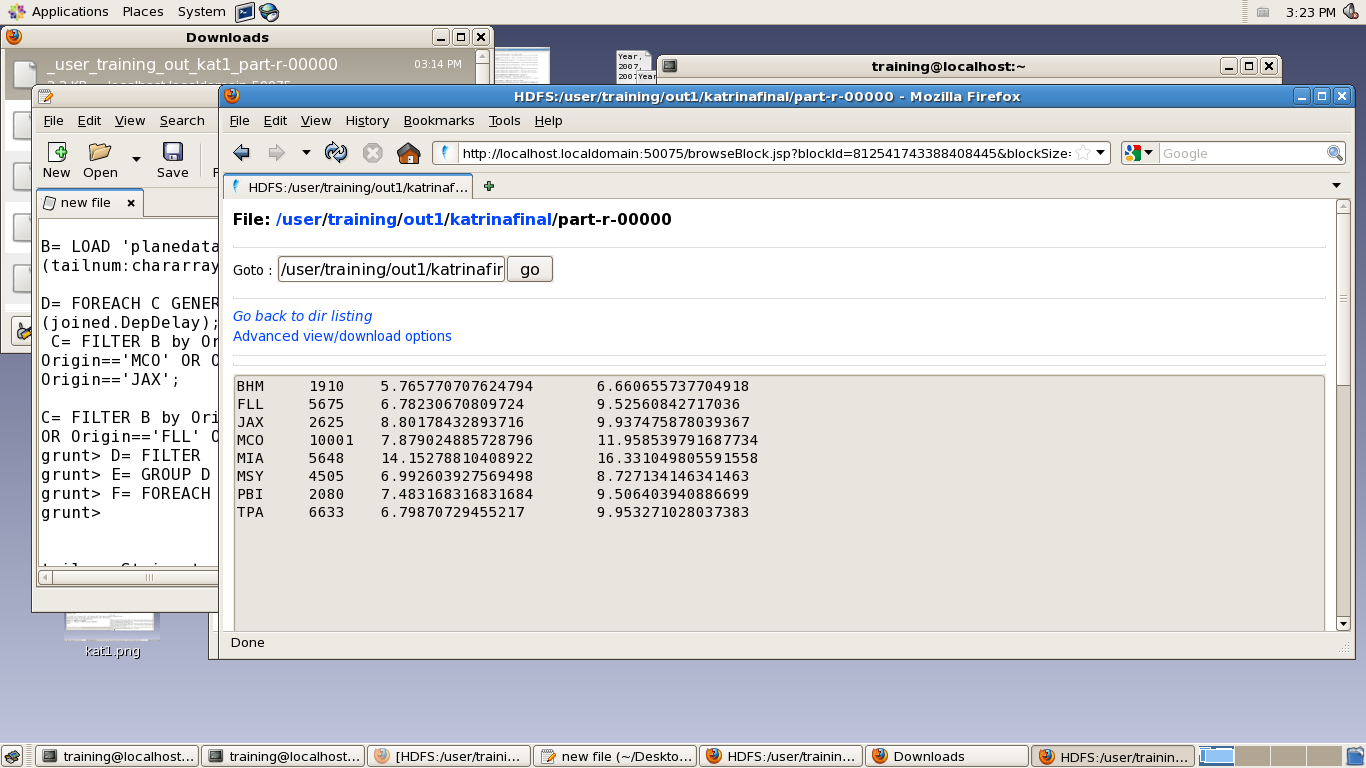
**ScreenShots**







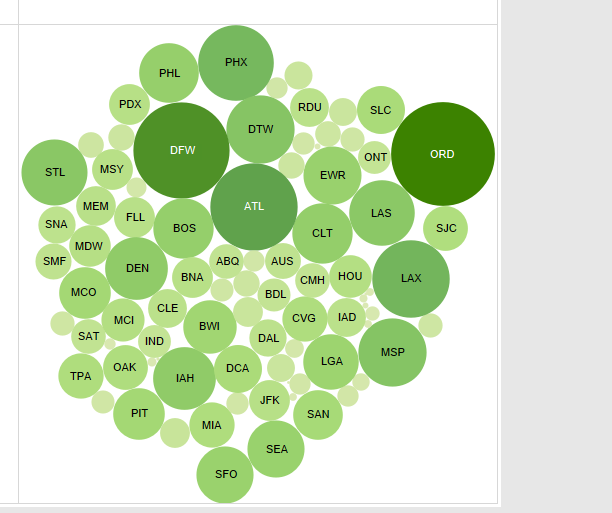




**Major Trendz in Analysis**

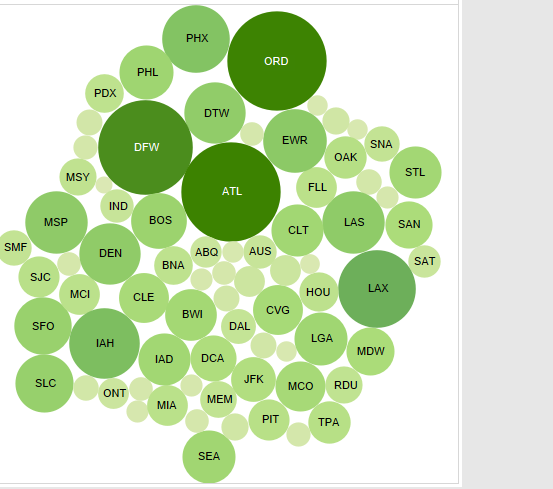
1. **Busiest Airport of the Year**

* Based on incoming and outgoing traffic(i.e. no of flights)

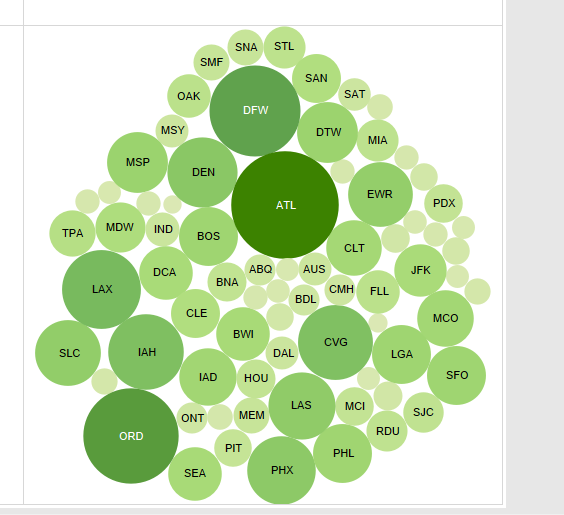


Traffic For year 2002

ATLANTA (ATL) , DALLAS(DFW) , O’Hairs Chicago (ORD) are the busiest Airports of 2002

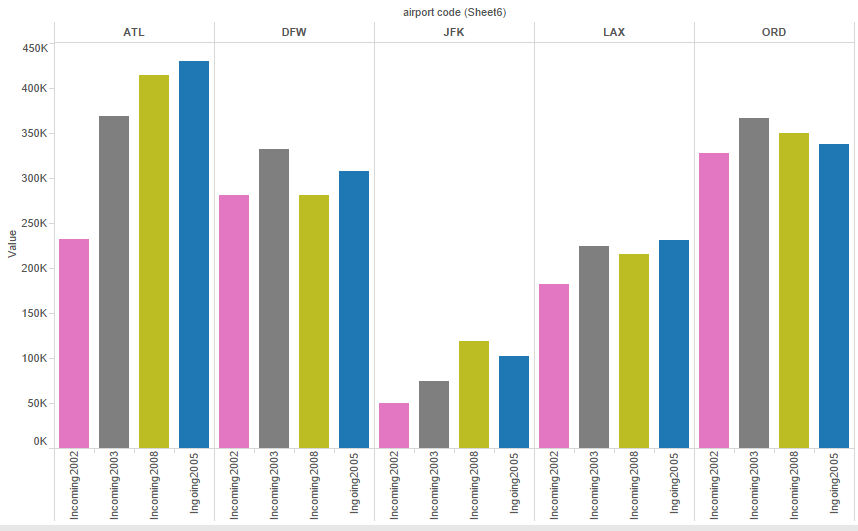


2003 -ATL brags no. 1 position in 2003

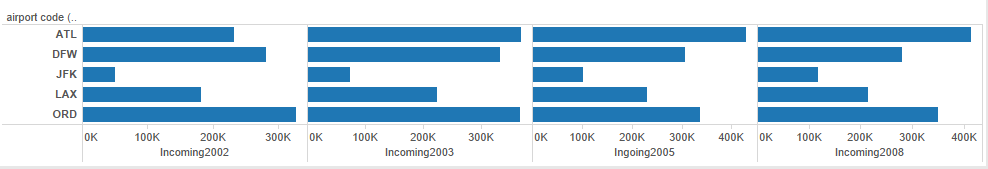


2005 - Atlanta Maintains the top Position

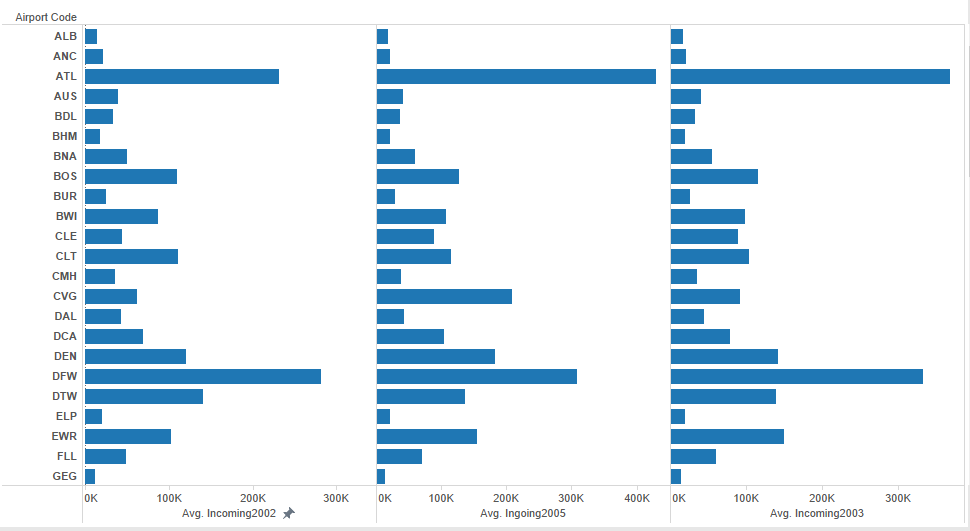
**2 .Growth of Busiest Airports of USA**



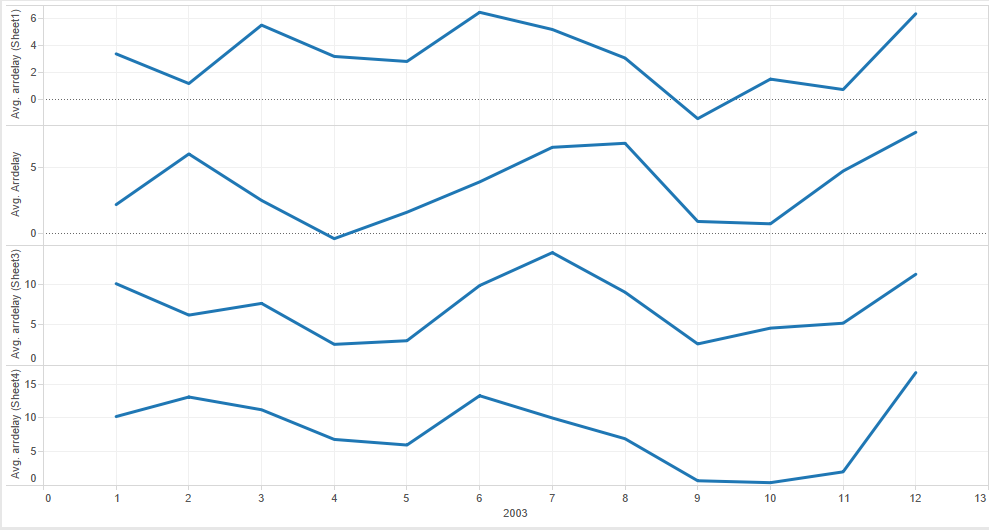
**3 . Comparison of Busiest Airports of USA**



**4 .Details of some Airports of USA**

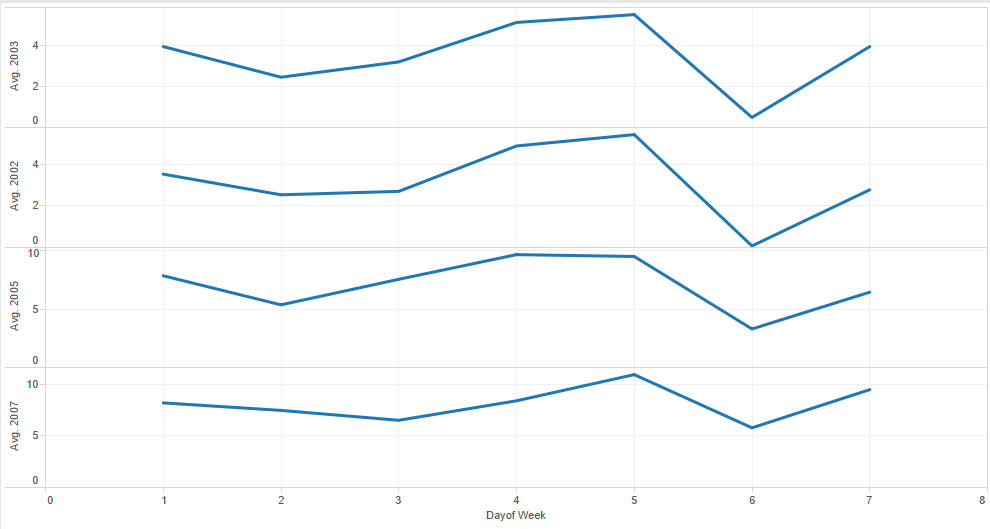


**5. Busiest Month of the Year - DECEMBER**

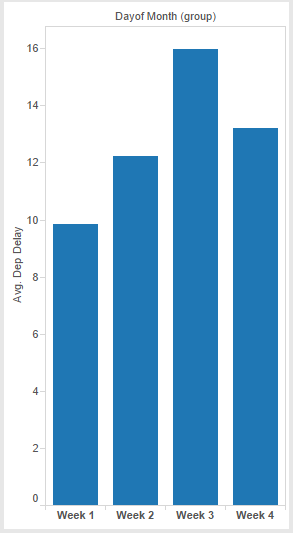


**Based on our analysis December is the busiest Month of year , while Sept and Oct have very low traffic**

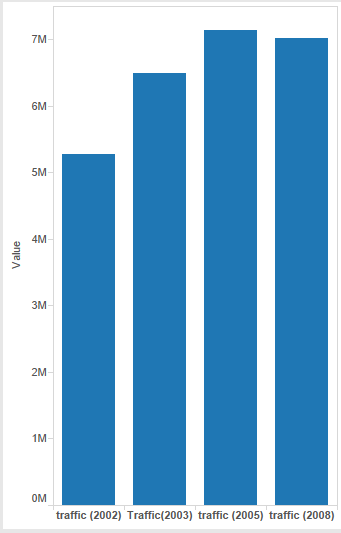
**6 .Best Day to Travel – Friday is the best day to travel as it has very less arrival and departure delay**



**7. Best Week To Travel**



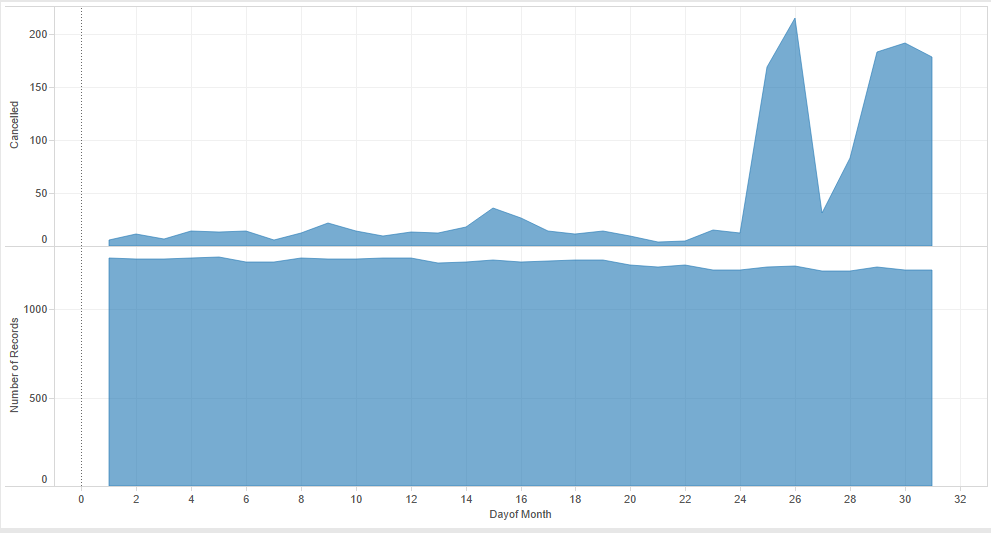
**8 Traffic per year** – shows no .flights took place during a year. In 2005 it crossed 7 Million ,first time in History



**Major Report During Katreena2005** - Analysis during the period of Katreena

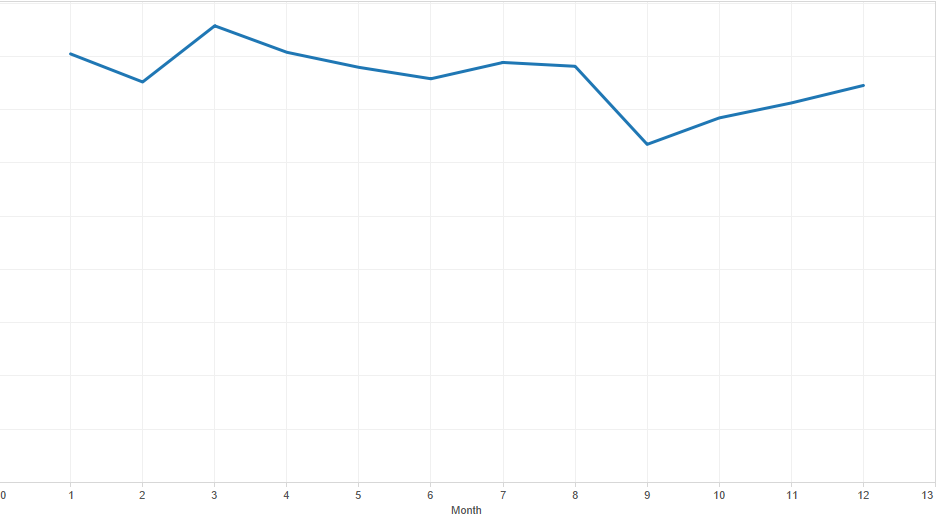
**Flights Booked vs Cancelled( Affected Regions )**

* **Katreenaoccured from 23rd Aug to 31st Aug**
* **Around 200 flights per day got cancelled during that period**
* **Whereas before 23rdavg was around 20 flights per day**



**Katreena Recovery**

* **AfterKatreena , traffic got reduced in affected areas ,we can see a sudden dip in month of September**
* **It took around 4 months to recover fully.**



**No. of Flights Cancelled –**

**Around 1500 flights per month got canceled during AUG SEP OCT**

