San José State
UNIVERSITY

Predicting Traffic Movements using Big Data Analytics

By
"IN SILICO"
Spring,2015

NAME	EMAIL
Adwait Kaley	adwait.kaley@sjsu.edu
Akshay Desai	akshay.desai@sjsu.edu
Apurva Patel	apurvaudaykumar.patel@sjsu.edu
Harsh Vyas	harsh.vyas@sjsu.edu

ABSTRACT

Predicting Traffic Movements using Big Data Analytics

By Adwait Kaley, Akshay Desai, Apurva Patel, Harsh Vyas

The City of San Jose Department of Transportation, has long been using state of the art infrastructure and latest technologies to deal with the constantly growing traffic. However, with the ever increasing traffic in the surrounding Bay Area, the problem is far from solved. The proximity to various sports stadiums, conference halls and convention centers, which attract crowds, does not help as it further leads to traffic bottlenecks and congestions.

Traffic Mismanagement also proves fatal for services like Emergency Medical Services, Fire Fighters, Police officials for whom time is of the essence. According to a recent survey, San Jose ranked 5th in the list of the top 100 most congested cities in the nation, a steep rise from last year's 13th position. It is of utmost importance for the traffic management authorities to deal with this problem at priority as it can lead to a chaotic situation in the city and surrounding areas as major highways in the state pass through the neighboring areas. They are concerned about this grave situation at hand as the current system would not be able to bear the ever increasing load of traffic.

This report aims to provide a solution for the traffic problem by harnessing the power of Big Data Analytics. The main cause of the problem at hand is management of the huge number of vehicles on road that continue to rise exponentially every year. Should the government invest money to build more roads instead of optimally utilizing the ones available? Absolutely not! Technology is revolutionizing and every imaginable way of reaching from point A to point B is possible. By using the data gathered from multiple sources, it is possible to do predictive analysis to manage, reduce and anticipate traffic movements at all times. The amount of data gathered from a variety of devices and will be huge and in disparate formats. We can collect, process and analysis this data by using Big Data Analytics to predict the traffic situation.

Cities are growing rapidly and mounting transportation challenges. That growth brings plenty of opportunities to help build intelligent transportation system. For Example, People use social media as a way to communicate with the world over the internet. The system can provide predictive data analysis by integrating and analyzing data from social media related to traffic in a particular area. This will help authorities to divert the traffic over to less crowded routes to prevent traffic congestions. This strategy can be immensely helpful for handling traffic during events which attract large crowds and prevent bottlenecks.

There are certain challenges in such data driven approaches. Management of the centralized data collection is one of the major challenge. That include how to collect real time data from the traffic to centralize and manage the data. Data can be collected using various technologies like GPS, CCTV, Sensors, etc. Data collected needs to be stored in databases which support handling of large volumes of data in different formats, which can be handled using modern day document based databases like MongoDB. Another challenge is the utilization of the collected data. Collected large amount of data can be processed and utilized to understand the behavior of the traffic flow also utilized for public security and criminal investigation purposes. Hadoop Distributed File System and Apache HBase can handle the challenge for storage. Hadoop allows real time data processing of the large amount of data.

TABLE OF CONTENTS

Chapter 1. Project Overview

Traffic is one the many menaces one has to deal with in daily life, and there is certainly no escape to it is no matter if you are the president or a common man. If the traffic is not managed efficiently, it would definitely be a chaotic situation throughout the city and this is something we cannot afford as time is if the essence to everybody. A single highway, if choked with traffic, could lead to affect various businesses and services alike for example, consider a city that relies heavily on the medical supplies reaching every day at the hospitals. If the medicines fail to reach the destination on time it may have catastrophic consequences. To tackle such situations, it is essential to leverage the latest technologies to attenuate the physical limitations of the transport department.

In recent times, the use of computers in managing the traffic systems has helped regularize the traffic to some extent by managing the traffic signal controlling. There exists many intelligent transport management systems nowadays. However, they are giving limited solutions in different ways , some are giving only traffic jam information or some specifically monitor speed indicators, some are giving routes based on the historical data. Our main aim is to introduce a user-friendly application which is capable of giving real time solution to the end user by considering the various parameters like, traffic jams, optimum route retrieval, travelling time, travelling speed, and weather conditions affecting traffic. Based on all this information from various devices spread out all over the city is estimated and the optimum route is made available to the user based on real time data and also certain predictions based on the past data collected.

The city of San Jose, experiences such problems on a regular basis. The authorities have tried their best to make use of the existing system to solve the recurring problems. As it is located right in the heart of Silicon Valley, owing to the presence of numerous corporates, attracts an increasing number of vehicle movement. The existing solutions are not fully self-sufficient to handle the traffic within and around the city. Also the frequent events organized in and around some major stadiums in the city lead to traffic congestions on certain days which turns out to be a traffic nightmare. Management of traffic thus, acts as a crucial entity that needs to be dealt with utmost importance as the number of vehicles on the road are going exponentially which the current system will not be able to negotiate for long.

We researched IBM’s InfoSphere Stream Big data Analytics tool which is capable of processing extremely high amount of data and concluding useful insights based on data processed. We aim to propose a solution which is based on predictive analysis and real time data monitoring. After collecting real time data, we shall be able to process it and also merge it with the prediction analysis made by the system for pattern matching. Also, combining this stream computing paradigm with Apache Hadoop to store semantically efficient data will help build a long term solution to the traffic situation. This will help us to find a user-friendly and most efficient solution in real time so that the traffic can be diverted to ease out the congestion in peak hour situations as well as in event of any special event which attracts crowds.

Chapter 2. Intelligent Transport Management System (ITMS)

It is essential to have a high quality transport system, which would make places easily reachable and people can stay connected. A transport system is said to efficient if it can manage new circumstances concerning safety, traffic congestion - by extracting useful information from the acquired data which is of benefit to the users and operators.

In United States, the government activities related to Intelligent Transport Management System, is driven by their focus on homeland security. It also includes the surveillance of roads via video cameras. It's most helpful during natural calamities to carry out mass evacuation of people and prevent loss of lives. Such a system has become a necessity in both developed as well as developing countries as motorization plays a critical role in their growth and development. Intelligent Transport Management technologies have a number of applications such as electronic tolling, parking guidance, speed limits, and real-time navigation.

2.1 Technologies used for ITMS

To implement different applications in ITMS such as car navigation system, weather information and speed monitoring, we need to make use of different technologies. Some of these technologies are listed below:

2.1.1 Wireless Technologies

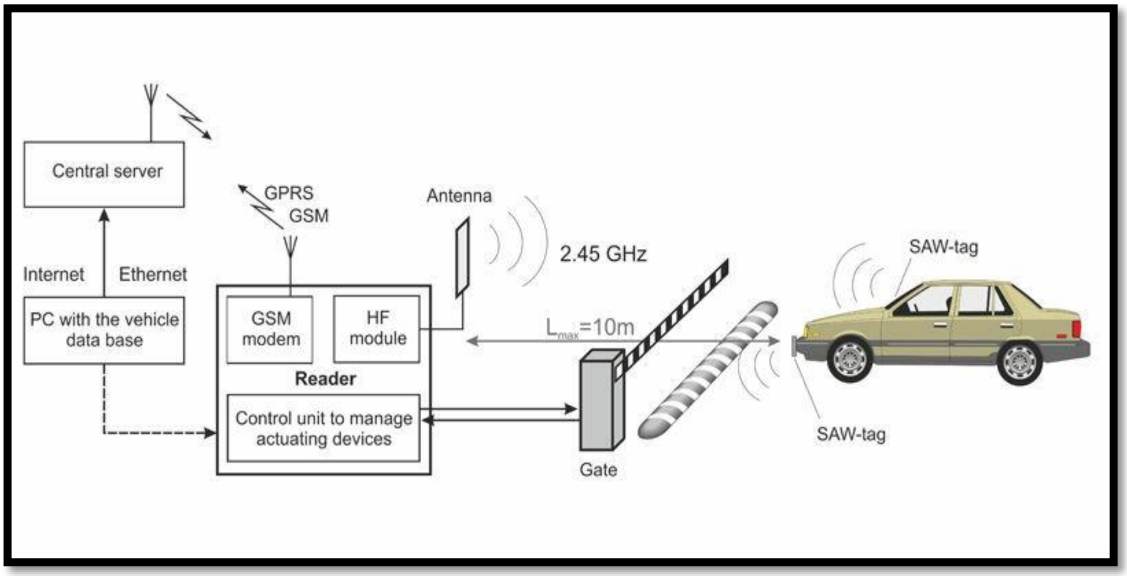
A number of wireless communication technologies have been proposed to be utilized in ITMS. Applications within ITMS for short and long ranges communication use Radio modem communication on UHF and VHF frequencies. Communications in the range of 350 meters are said to be short range communications that can be achieved by using IEEE 802.11 protocols. Wave and Dedicated Short Range Communications are the ones which are recommended by ITS of America and United State Department of Transportation. Though these networks are short, we can theoretically extend them using mobile ad hoc networking or mesh networking.

On the other hand for the long range communications, there are some suggested infrastructure networks such as WiMax (IEEE 802.16), GSM, 3G. But these types of infrastructures need extensive and expensive setup. The eCall and behavioral tracking functionalities need to be supported by ad hoc solutions in the form of Techmatics 2.0., especially used by auto insurance companies.

2.1.2 Sensing Technologies

Sensing technologies has brought an inexpensive and intelligent solution to the table. Advancements in the fields of telecommunications and information technology such as RFID and beacon sensing technologies have been a great support for ITMS. They are mostly vehicle and infrastructure based networked systems. Apart from being inexpensive they are also less prone to damages, for example in-road reflectors or devices that are embedded in or around the roads, buildings, post and signs. They could be manually dismantled as and when needed. Video automatic number plate recognition and vehicle magnetic signal detection technologies are some of the applications of sensor technologies which could be placed at appropriate intervals to increase sustained monitoring of vehicles moving in critical areas.

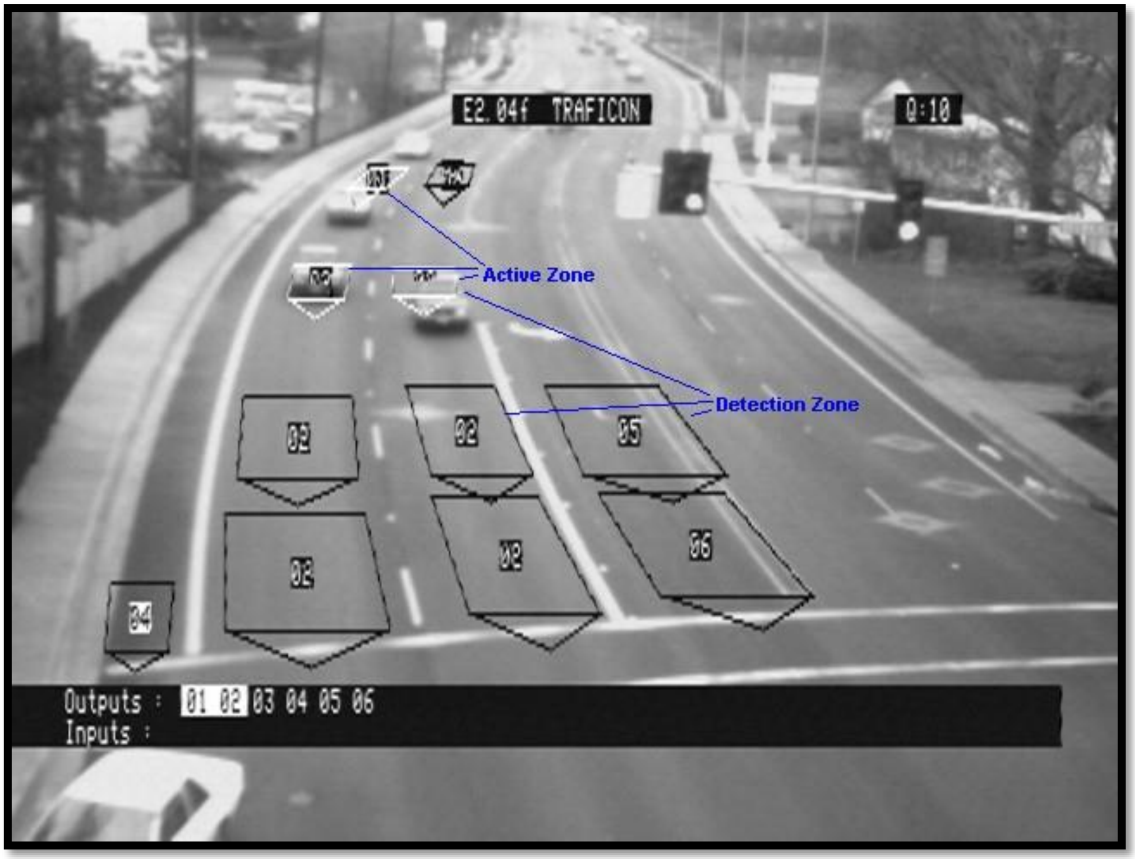
Another form of sensing technology is the Bluetooth detection, yet another inexpensive technology. Bluetooth is a wireless form of communication usually used to connect phones, computers headsets etc. Bluetooth sensors placed on the roads can connect to the Bluetooth on the vehicles to detect the MAC Address. If they get connected, then we can get a lot of info such as calculating travel time and also provide data like origin and destination coordinates.



RFID Parking System

2.1.3 Visual Technologies

Use of Video cameras in vehicle detection such as traffic flow measurement and auto incident detection has been crucial. Inputs from video camera are fed into processing units to analyze the changing characteristics of the video image. Here the video cameras are not installed in the roads, hence it is known as “non-intrusive method of traffic detection. They are attached to the poles high above the ground level or structures above or adjacent to the road. With the placement we also need to make sure of the basic standard image which needs to be referred to. This standard set may contain lane line distances or at what height is the camera placed above the road level. The processing units take input images from a number of cameras and process all that data to give us necessary information. This information can be vehicle speeds with respect to the lane, count of vehicles on the road or particular lane. Some advanced systems also provide additional information such as wrong turn alert, spotting stopped vehicles and gaps between vehicles. The figure below shows how detection takes place with respect to the lanes, which cars are in motion and their likely path.



Video Camera

Chapter 3. The Big Data Solution

Our main discussion in this chapter revolves around how big data helps in evolution of computing specially for real time applications.

3.1 Big Data Analytics

What is Big Data?

In simple language we can say Big Data is to derive new insights, new ways from previously known data and use those insights and ways to improve your business etc. Technically, Big Data is a wide term for information sets so expansive or complex that customary information handling applications are deficient to process them.

Why should we use it?

A simple answer would be, to learn from the historical data and identify patterns so that if we encounter a similar situation in future we could devise a strategy beforehand to work upon.

Just to take an overview of the Big Data, we define Big Data to be a V3(Vcubed) paradigm. i.e. **Volume, Variety, Velocity, Veracity(optional)**

- **Volume**: It's a bulky data coming from different sources. No matter what source it is but it is a huge amount of data to handle
- **Variety**: The data coming from different sources is the variety of data. For example in our system data will be coming from a video source or GPS signals or from sensors or from any device which measures the weather factors.
- **Velocity**: The data may be coming in huge amount within fraction of seconds, may be it can come within minutes depends upon the need for the application.
- **Veracity**: Truthfulness of the collected data from which we can derive different insights. So fourth factor is also important to the application as if it is equally useful to derive solutions for the application.

As our main focus is to use IBM InfoSphere Streams software to handle Big Data. This software has very good features for handling big data that we want to present. They are as follows.

1. Meta Data, business glossary, and policy management
2. Information integration
3. Data quality
4. Master data management(MDM)
5. Data life cycle management
6. Data privacy and Security

We have a separate portion for IBM InfoSphere Streams understanding in the following chapters. There are some features that makes IBM platform to be independent in the BIG Data filed and that are as below.

- It has the broadest platform to integrate the data and governance in the market.
- With IBM you can find a technology that is a low risk and for the client's usefulness perspective
- The software can be used for the faster deployment purpose and more focused on agile development lifecycle.
- They had prebuild knowledge rules and methods based on their past experience which is very useful for deriving the solutions.

3.2 Availability vs Efficiency

There are three basic factors which affect the product completion: Availability, Efficiency and Productivity. Amongst these three, Availability and efficiency are most effective for the product to be in the market. We are handling Big Data so we need to take care of this two factors at priority.

Availability:

The proposed application operates in real time situations so data needs to be accurate in real time. Availability in our project is for the data which needs to be accurate as well as we should be able to use it whenever it is needed. We are using Info Sphere Stream software that is very useful to make any real time application for big business. We choose that software because of its effectiveness in handling huge amount of data which is highly available.

The data is coming from different sources like video cameras mounted on signals or cross roads, data coming from GPS and from the sensors mounted across the roads. These are the variety of data and needs to be handled accurately. IBM info sphere Stream has this functionality which is very useful in handling big data with efficiency.

Efficiency:

As availability is the major factor affecting the application, efficiency is also equally important. The term itself means how productive or how effective our application is in terms of application usage perspective. Efficiency in relation with the availability is important. We are drawing insights from the available data collected from different sources. This data is going to be used to tell the user about current traffic conditions nearby, about traffic speed, about most optimum route to travel from one place to another place, about how the weather conditions going to affect the traffic conditions. These are the data related to different sources. These data are collected and then analyzed and based on that we are going to make decisions for the optimum routes. The whole process should be highly efficient, user should not feel uncomfortable about or feel confused about the application data. This process is taken care by our IBM Info Sphere Streams.

3.3 Risk vs Efficiency

Risk:

Risk is one of the factors that is affecting the products' usefulness.

We can list out three risk factors affecting the big data environment and our application.

1. Requirement Satisfaction
2. Business model changes
3. Pricing factor

Above risk factors are common in any application. Requirements should be clear and should be achievable. We have one problem of handling Big Data and that problem is solved by IBM InfoSphere Streams. Whenever it is required we are going to change the respective Business model for the system's efficiency. Our system will be free of cost as if it is for the use of common people around.

Efficiency:

We are considering Risk then we need to consider efficiency of the product with the Risk factor too. The requirement satisfaction risk is directly in relation with the Efficiency factor. If we want to satisfy every requirement then it should be highly efficient as well. If we want to change our business model then it should not affect the product's efficiency. Efficiency comes with a price that's why pricing factor is directly related to the pricing factor. Our application will be free to use so we are not worrying about pricing factor right now.

Chapter 4. Data management

In this chapter we aim to show the data management techniques that we plan to use in the core system. We are going to list certain technologies that we took into consideration for Data collections and data monitoring and tools for that.

4.1 Data Collection Technologies

Collection of huge amount of data is a big challenge for a system. There are many sources from which data can be gathered. Data collection technologies should be highly optimum and less costly to compete with the system performance. Our data can be digital or analog. Data can come from an Audio/Video source or it can come from any sensors, GPS, any device which can transfer data. Our main challenge is to collect the data from various sources and to integrate them and process them to find out a real time solution.

In our System we are dealing with digital signals, GPS data, Video/Audio data processing and there are various ways of handling those data using disparate technologies.

Digital Signal Processing (DSP)

Data coming from the sensors mounted on roads side and /or any vehicle is in the analog form. Digital Signal Processing is a mathematical way of manipulating the information coming from the signals. Main goal of DSP is to measure, filter and compress the real world analog signals. The initial step to handle the signals is to convert them into digital form which requires an Analog to Digital Convertor. ADC will convert the signals into streams of digital values. We are using IBM's InfoSphere Streams to handle the data and this software works on stream computing techniques. Stream computing continuously handles real time data and analyzes it to find useful data sets. Info Sphere Streams can handle data from different sources and some sources are sensors and radars. Sensors and Radars send data in Analog form. So we need Digital Signal Processing to handle these data.

GPS Data Processing:

We are using GPS technology to keep track of the vehicle and to help the user find out the optimum way in the current traffic situations. So we need GPS data. GPS has its own format. We can handle GPS data using GPX or GPS Exchange Format which is an xml design developed to handle GPS data for any Software Application. This data can be used to keep track of routes and the format is an open source. The tags in the xml format stores data for time, location, elevation and can be used to interchange data between the GPS and the software application. This will help us store the data in xml format and to process it through Info Sphere stream and then give it back to the user in the GPS device to locate the optimum way of travelling.

Video and Image Processing:

Big data is not something which has emerged just yesterday; it is one of the rapidly growing technologies in IT. 80% data which is collected is unstructured data such as video or photo. If you just consider the handheld devices, they produce millions of pixels of data. Similarly large sets of data such as images and videos(which as stream of images) are also generated from hundreds of cameras that are placed at highways, road, parking lots and cameras around office areas. These data sets are large because they provide continuous stream of data at real-time. This huge real time data is the raw data. This raw data is analyzed and intelligent information is extracted. This big data will enable us to identify patterns and get to an accurate solution for our problem.

4.2 Data Storage Tools

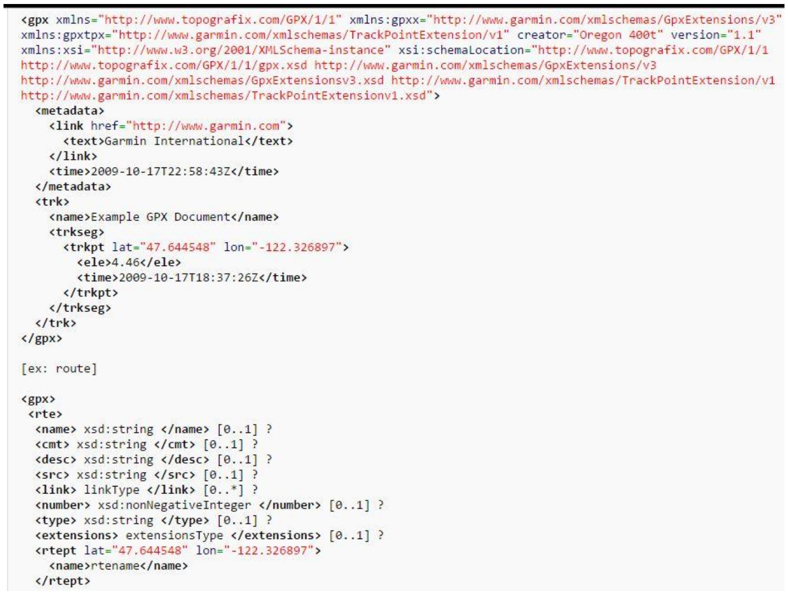
As we are dealing with different types of data and we need to store these data in storage devices. We are going to need advanced processors like Intel Xeon processors that have tremendous amount of processing speed especially for the video and audio files. It processes around 500 kb digital image with the speed of 250 times/second. The data storage capacity should be more because we are dealing with huge amount of data. So sufficient amount of storage capacity clusters are needed to store the data. This clusters can be mounted on different places to collect the data.

4.2.1 Data Monitoring

There are many software available in the market for monitoring data. We are dealing with analog signals, GPS data and video/image data. We can have different monitoring systems to manage all the data. Let us discuss some of them.

GPS Data Monitoring:

GPS data monitoring automatically detects the location of the GPS device and handle it in the network mode. This data can be stored and can be analyzed afterwards for the required results. Tools that can monitor, analyze and chart data are 32 NAVSTAR and 24 GLONASS satellites. There are 56 satellites. The data exchange format that we saw in data collection technologies is GPSX and using that data collection technology it is possible to store the data and monitor it. Here is one format conversion of GPS data into an XML format.



```
<gpx xmlns="http://www.topografix.com/GPX/1/1" xmlns:gpxx="http://www.garmin.com/xmlschemas/GpxExtensions/v3"
xmlns:gpxtpx="http://www.garmin.com/xmlschemas/TrackPointExtension/v1" creator="Oregon 400t" version="1.1"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://www.topografix.com/GPX/1/1
http://www.topografix.com/GPX/1/1/gpx.xsd http://www.garmin.com/xmlschemas/GpxExtensions/v3
http://www.garmin.com/xmlschemas/GpxExtensionsv3.xsd http://www.garmin.com/xmlschemas/TrackPointExtension/v1
http://www.garmin.com/xmlschemas/TrackPointExtensionv1.xsd">
  <metadata>
    <link href="http://www.garmin.com">
      <text>Garmin International</text>
    </link>
    <time>2009-10-17T22:58:43Z</time>
  </metadata>
  <trk>
    <name>Example GPX Document</name>
    <trkseg>
      <trkpt lat="47.644548" lon="-122.326897">
        <ele>4.46</ele>
        <time>2009-10-17T18:37:26Z</time>
      </trkpt>
    </trkseg>
  </trk>
</gpx>

[ex: route]

<gpx>
 <rte>
  <name> xsd:string </name> [0..1] ?
  <cmt> xsd:string </cmt> [0..1] ?
  <desc> xsd:string </desc> [0..1] ?
  <src> xsd:string </src> [0..1] ?
  <link> linkType </link> [0..*] ?
  <number> xsd:nonNegativeInteger </number> [0..1] ?
  <type> xsd:string </type> [0..1] ?
  <extensions> extensionsType </extensions> [0..1] ?
  <rtept lat="47.644548" lon="-122.326897">
    <name>rtename</name>
  </rtept>
```

Fig 4.1 GPSX Format

As the data is converted to xml format this data can be analyzed easily. We can figure out data realted to time and location easily. This data is analyzed by IBM info Sphere stream and it is going to pruduce a useful output from it.

Video Data monitoring:

IBM itself has one video data monitoring software which does video analytics to identify patterns , attributes, events .This video analytics software has the ability to analyze the data and automatically generates the alerts and facilitates the forensic data to identify the patterns , trends , incidents and to use the data in the way you want to use the data.

This software has many different features that can be used in any video analytics application.

- **Rich content-based indexing and advance search capability** to find out specific near real time historic event.
- It has **Safety, Security, Investigation facilities**.
- **Open standards and extensible architecture** that we can extend to use in our system.
- **Integration with other software.**

We require these features to monitor our video data coming from the cameras. The data should be clear and should be real time. These technologies can help us find out congestion at a particular area. We are going to monitor video data and supply our data to IBM infoSphere Streams which is going to make decisions and going to provide the user the best possible answer about it.

Chapter 5. IBM InfoSphere Streams

IBM introduced a product named IBM InfoSphere Streams in April 2009. Basic concept of Stream is to provide a facility to the clients to process the massive volume of data with an extreme speed so that client can analyze the data for the improvement of business decisions.

IBM research team almost worked for over a decade to improve the computation technologies to process large volume of data and to analyze the data as fast as possible. How their work is important to the world? Consider how the traffic prediction can be benefited by this technology. Large amount of data such as, GPS data and CCTV video images can be processed for the prediction of the traffic. Also, at the stock market where large info streams are processed per second and it keeps on changing. Imagine how much time, resource and money can be saved if we predict the critical situations before it occurs just by processing the real time data.

5.1 Overview

To understand the InfoSphere it is necessary to have understanding of what a Stream is. Our purpose here is to introduce the concept of Stream and how it was designed and implemented for the real world data.

Amount of data available to the enterprise and other business organization is increasing day by day as the technology growing rapidly it helps collecting more and more data. Every company is trying to collect this data and convert it into useful information which can help them expand their business. To achieve this goal it is necessary to have a system which can process and analyze large volume of data instantaneously without the storing it. Stream provide a platform where user can develop application and it provide a runtime environment where user can deploy and run the application to analyze the data. Streams also allow the dynamic modification of an application. The data streams can be collected from the sensors, cameras, GPS devices and it can be in the form of text, image, video, audio or radar input.

Application can be developed in InfoStream Studio and it can be deployed on the Stream Runtime environment. By Streams live graph performance can be monitored. IDE of the Stream is based on an open source Eclipse project. Developer can specify the operation they want to perform on the stream such as filtering, merging, transforming and performing complex mathematical functions. Stream studio is not only used for development but it can also be used for the debugging. Debugger also support the runtime environment and the stream live graphs. User can inject and inspect the data in the developed application. Compiler optimize the performance

5.2 Architecture

InfoSphere architecture explain the execution and the development of the application and It shows how the large data stream is processed by the system. It also explain the functionality of the system. InfoStream address many objectives. It provides dynamic adaptability and scalability. Load balancing, scheduling and high availability is a significant functionality of Stream.

5.2.1 Concepts and Terms

Computing system is significantly changed by InfoSphere Stream architecture. Some application processes need an environment for continuous data stream processing. Streams provide development, runtime and debugging platform to the applications.

5.2.2 Component Overview

InfoSphere has integration of command line tools as well as graphical tools. Also it has compatibility with the tools providing monitoring, administrative and analytical features. It is also providing runtime, development and debugging tools.

5.2.3 Runtime System

Runtime environment is made of components and different services. It integrates with the operation system. This integration provides high performance. Runtime environment is an execution environment where developed application can run. Environment itself automatically organize the resources and it monitor the state change of an application.

SPL Components:

InfoSphere has a language called Streams Processing Language (SPL). SPL allows developer to create an application without the basic understanding of the Stream. It has many built in functions to import the data outside the Stream and export the processed data out of the Stream. Join and aggregation function is also supported by SPL.

Underlying runtime system creates a data flow graph on deployment of an application. When new data streams are submitted to the application, the runtime system provides resources to the new data without disturbing the operation of the existing data and runtime system continually monitors the state of the application and the data streams.

Results can be export outside the current system. Runtime system sink TCP operations to send the data streams out of the system where it can be used by different system.

SPL has following components:

Tuples:

Tuple is a piece of data which contain attributes or variables. Generally, data in the tuple represents a state of any application at a particular time.

Projects:

It contains some of the configuration settings or the files related to the Stream. It is kind of a container for the streams.

Applications:

Application is an implementation of the concept which runs on the runtime system to perform the operation. It process the stream of data.

Operators:

Operators are building function or the blocks of SPL. Operators can be created according to the need of an application. Example of some basic operators are join, merge, sort, filter etc.

Runtime Components:

Runtime system is made of services and components. Some of the runtime system components are as following.

- **Interface:**

  Interface is a set of interacting components runs on different platforms. More than one instances can be created for an installed application. Instance are independent.

- **Host:**

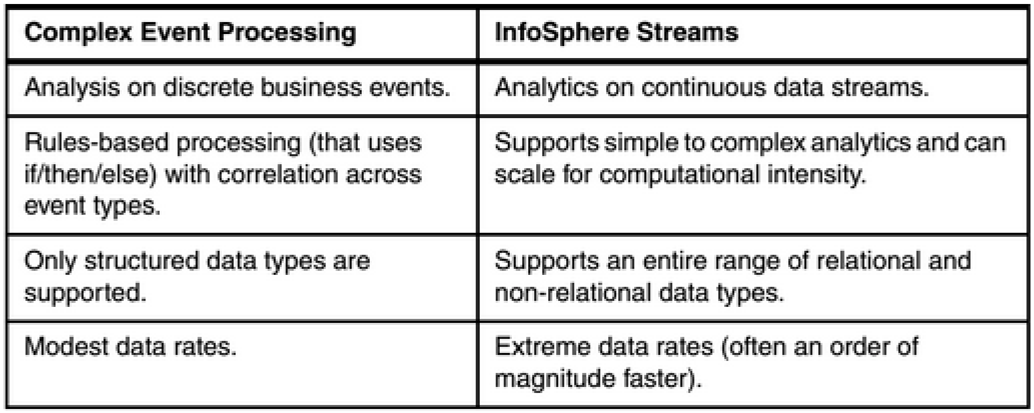
  Host is similar as an operating system and also known as node of the system. Multiple hosts can exist for a system such as management host and application host.

- **Host Controller:**

  It runs on every application service and manages the request made by Streams Application Manager. It decides when to start and stop the service. It monitor the process elements.

5.3 Comparison between InfoSphere Streams and CEP

Complex Event Processing is a new business concept for event processing. It identifies a characteristic from task of processing and business events. Streams features are somewhat similar to CEP but Stream supports high data volume. CEP processes hundreds of thousands of message per second whereas Stream can process millions of message per second. CEP can handle discrete events, Streams can handle discrete events as well as real time data such as video and audio.

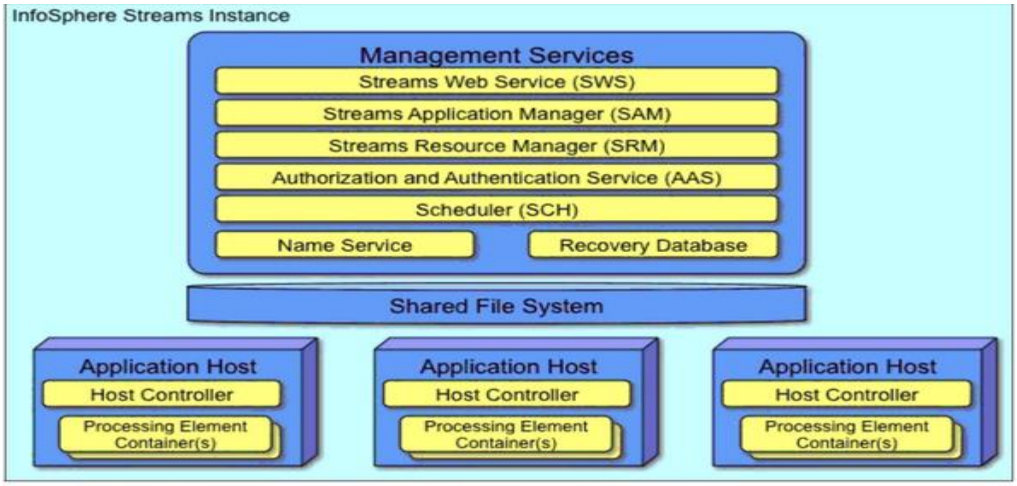


Complex Event Processing	InfoSphere Streams
Analysis on discrete business events.	Analytics on continuous data streams.
Rules-based processing (that uses if/then/else) with correlation across event types.	Supports simple to complex analytics and can scale for computational intensity.
Only structured data types are supported.	Supports an entire range of relational and non-relational data types.
Modest data rates.	Extreme data rates (often an order of magnitude faster).

5.4 Deployment

5.4.1 Runtime Architecture

InfoSphere streams provide a runtime engine that administers, configures, runs and streams the data. Streams instance is a software configuration which runs on one or more servers. This configuration provides the environment to manage, start, stop or monitor the streams. Streams instances are service processes that manage and run the application. Stream instances are independent and they operate independently even though they share some of the resources.


InfoSphere Streams Instance
Management Services
Streams Web Service (SWS)
Streams Application Manager (SAM)
Streams Resource Manager (SRM)
Authorization and Authentication Service (AAS)
Scheduler (SCH)
Name Service
Recovery Database
Shared File System
Application Host
Host Controller
Processing Element Container(s)
Application Host
Host Controller
Processing Element Container(s)
Application Host
Host Controller
Processing Element Container(s)

- **Management Services**

Streams instances can host on a single server or multiple servers. If the Stream instance is on a single server then the service is stored on the same server and if the instances are in different servers then the services are spread across the different servers. Some of the management services are as following:

- **Streams Application manager (SAM)**

SAM manage the application in stream runtime system. It manage the job management task such as job submission job cancellation. SAM interact with the host controller and scheduler to process the application.

- **Streams Resource manager (SRM)**

SRM initializes the stream instance and it monitor all the instances. It collects performance metrics necessary for scheduling and administration of the system. It interact with the host controller to get the information about the performance metrics.

- **Scheduler (SCH)**

Scheduler is a service which make placement decision for the application on the runtime system. SCH interact with SAM to get the job deployment information and it interact with the SRM to get host information on with the service is going to be run. SCH also get information about runtime metrics from the SRM.

- **Name Service (NSR)**

NSR is a service that store a reference for each service and all the component of the given instances. It provide reference to each instance component so that the component can communicate with each other by the name reference.

- **Authentication and Authorization Service (AAS)**

AAS authorizes and authenticates the operation of an instance.

- **Streams Web Service**

It is a service which provide web access to the instances of an application. But SWS must be running on the host for the web access.

- **Recovery Database**

If the recovery setting is setup then it recover the state of an instance. It records the state and reaches to the state when application fails and when necessary.

Application Host Service:

Application host service run on each host server:

- **Host controller**

Host controller runs on every application host service. Host controller manage all the job submission request from SAM such as start, stop, monitor, PE. It collect performance metrics from processing element and submit it to SRM.

- **Processing Element Container (PEC)**

SPL consists of many operation. This operation groups together at the compile time and this partition executes in a container called PEC. Each partition run by a PE. PEC is monitored by host controller.

5.4.2 Streams Instances

Stream instance work as an autonomous element and it can be configured by multiple options. These options are mentioned in stream tool man command in IBM InfoSphere. New instances for a stream application can be created by Instance manager. Instance manager is a new service to manage the stream instances. Stream console can be used to change the property of an instance.

Stream instances are isolated from each other. We can create multiple stream instance on same Stream installation but this option is not a feasible option. Using multiple instances on a single physical resource may cause contention of a physical resource.

Stream Instance ID:

A unique ID is assigned to each stream instance which differentiate each instance from each other. Format of stream instance ID is ‘instance-name@userid’. There are certain requirements that should be followed in instance name and userID.

Stream Instance Configuration:

To use multiple instances and to configure the instance some basic concepts should be understood that how a instance should be categorized. Stream Instance configuration is done by usage pattern and configuration or purpose.

Stream Instance File:

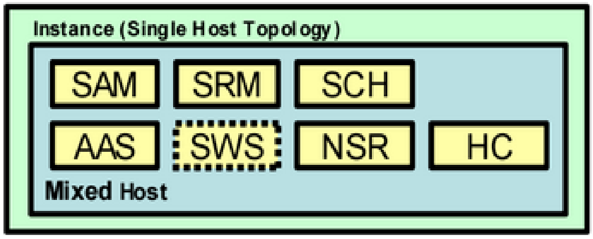
Apart from running an application and creating stream instances some of the component also matters are file system based components. Instance of a stream interact with log files, configuration and user authentication files.

5.4.3 Deployment Topologies

Deployment topology is a part of a planning phase of an application. Topology affects the performance of an application. Topology is chosen based on the need of an application. There are three basic topologies for deployment of an application:

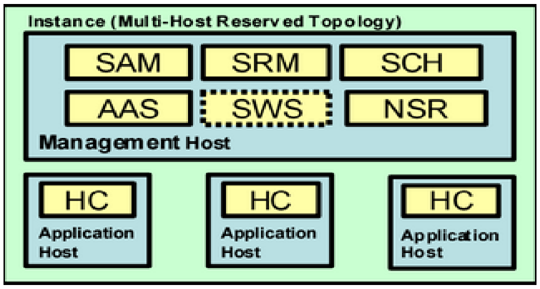
Single Host Topology:

It is a single host computer which run a single stream instance. It is a simplest topology. This topology is useful for low volume application and for learning streams and developing applications.



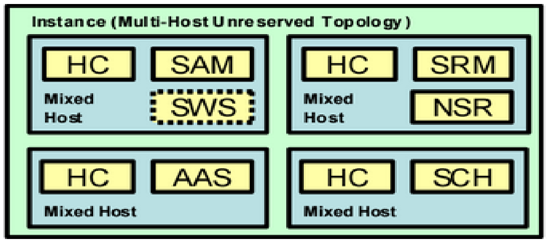
Multi-Host Reserved Topology:

For production environment, integration and test this is the most recommended topology. In this topology the instance management services are run by Management host which is known as a reserved service. And application services are run on the Application hosts. There are multiple application host on this topology.



Multi-Host Unreserved Topology:

In this topology the management and application services can run on multiple node. So the services are spread across the nodes. Most of the node have a mixed host. In some of the large cluster application there are application host in the topology.



Chapter 6. Real-Time Traffic Data Handling with InfoSphere Streams

6.1 Real-Time Traffic Data Handling with InfoSphere Streams Stream Analysis

Real-time device monitoring with IBM Info Sphere Streams high velocity stream processing will be capable of capturing data from various sources which include sources like Traffic Signals, Traffic Cameras, etc. on real time basis. This captured data will then go through high speed in-memory processing cycles to perform on the fly pattern matching. Through this mechanism, it will be able perform analysis in real time and simultaneously carry out predictive analysis along with at the moment statistics of similar components. There is a need to use the real time analysis methodology Real Time Analytic Processing (RTAP) used by IBM Info Streams. RTAP focuses on taking the proven analytics established in OLAP to the next level. Data in motion and unstructured data might be able to provide actual data where OLAP had to settle for assumptions and hunches. The speed of RTAP allows for taking immediate action instead of simply making recommendations.

IBM InfoSphere Streams takes a fundamentally different approach to continuous processing and differentiates itself from the rest with its distributed runtime platform, programming model, and tools for developing continuous processing applications. This gives the users the flexibility to modify the contents at any later point.

This is a critical aspect as the demands of the application to process traffic movements as every bit of information is of utmost importance.

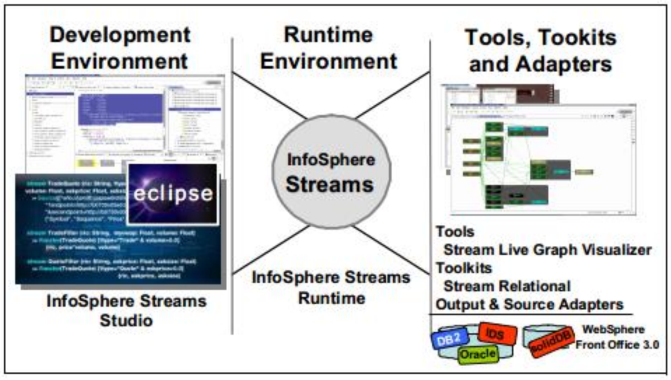


Fig.6.1. Components of IBM Info Sphere Streams

6.1.1 Implementation

The basic idea behind implementation of this solution with IBM InfoStreams is to analyze the parallel incoming stream from thousands of sources like sensors, traffic cameras, etc. and carry out analysis with the Stream Analysis engine to identify potential aspects that could occur based on the inputs. For notification and predictive analysis, the solution will make use of existing input feeds that currently go to the monitoring tools. This data will be fed to Apache Hadoop, which will make use of the specific algorithm and store semantically intelligent data. The data will be stored in different clusters, which are closely related and weighted upon by the computation algorithm based on routes. If the incoming stream contains specific routes retrieval criteria, the scenario would be considered as an indication of possible match based on the weighted average and a proactive computed solution will be fetched from the clusters based on the closest match and sent to user as the closest matching solution.

Analytical Traffic Congestion Notification will be another prime feature of this solution. Apart from utilizing the incoming feeds about traffic statistics for route prediction, this data will also be used to perform a parallel search across the data to gain insight about all the routes associated in the surrounding area of similar kind and provide a quick performance snapshot of all the routes at a given instance. Analytical Traffic Congestion Notification will act as a source of information for the transportation authorities to monitor the traffic situation of a route at any given point of time, also alerting the officials of any sudden irregularities observed than normal flow. This will help narrow down on determining the root cause of the congestion at the earliest. Furthermore, in event of such a situation the system will divert the traffic by feeding the next optimum route if the congested route is requested. This will allow authorities, the flexibility and time needed to handle the deviation.

6.1.2 Incremental Processing

The data received by the Streams Engine is of a very large volume and hence if we manage and try to store this data into a relational database. There would be a tremendous load on the Database Management Systems to handle such a load of data and may lead to inconsistencies. In addition to it, the large processing time needed to query the data will be a significant factor that could slow down the entire process of getting real time data analyzed and provide efficient routing to the user. Thus, we need to make use of incremental processing of data and eliminate the need to store the entire data set and process the data directly from the streams provided by the Streams Engine. Standard database servers generally have a static data model and data, and dynamic (albeit often long running) queries. IBM InfoSphere Streams supports a widely dynamic data model and data, and the Streams version of a query runs continuously without change. Queries in IBM InfoSphere Streams are defined by the Streams Application, and run continuously (or at least until someone cancels them).

As Streams Applications are always running, and continually sending selected data that meets the defined criteria to the consumers, they are well suited for answering *always-on* or *continuous* questions.

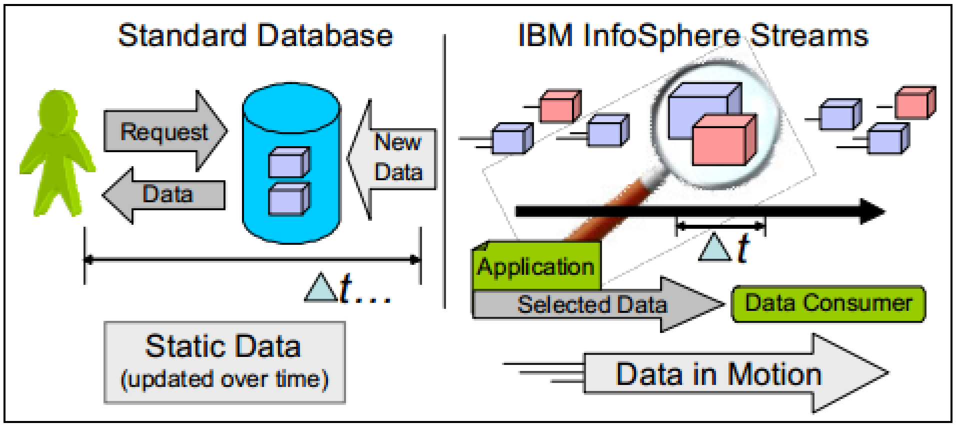


Fig.6.1. Comparison of Standard Relational Database to IBM InfoSphere Streams

6.1.3 Suggested Solutions

To manage the traffic problem, we suggest that the application to be developed leverage the key features of the IBM InfoSphere Streams, mainly for processing of data and use Apache Hadoop to store intelligent information to derive patterns from. Thus, we would be able to successfully use the benefits of both the independent entities and would be easier to manage them. IBM InfoSphere Streams, has inbuilt algorithms which manages the data streams into appropriate datasets to process real time data. As the traffic data is to be processed from variety of sources it is quintessential to have a stream software processing that is able to handle the data. Streams has two important aspects, a runtime environment, which includes platform services and a programming model, which determines the best way to service the request.

Also, the data to be stored can be stored using Apache Hadoop on a group of clustered servers, this is mainly because of the flexibility offered by the MapReduce, to efficiently store data across the servers.

6.1.4 Solution Aftereffects

Semantic Stream computing stored in collections of traffic related data to identify patterns and set up trends by which routes could be classified as congested or free. Graph based analysis for decision making will be generated to bolster the process of setting up new infrastructure or expanding the existing one. Device specific, vendor specific and function / operation specific graph based trend analysis will be one of the prime outputs of this solution. Thus, an efficient solution can be provided to handle the traffic management efficiently with the help of Big Data Analytics.

Chapter 7. Case Studies

To get ideas for our project we researched certain topics with the help of internet and found some exciting systems and researches ongoing on Intelligent Traffic System. These systems are giving a convenient solution to manage traffic in particular cities.

7.1 InfoSphere Streams(Dublin City):

Dublin city is growing much faster. Mainly, due of its fast growing industries and also due to the population growth there. Now the problem was to manage the increasing traffic. The Councils' Traffic Control Center started working with local transport operators to manage bus routes and cars and all types of commutes. The city council has many gadgets and instruments to invest for the system that would have helped them in the intelligent traffic system.

This is the software we are using as for the support to the Big Data solutions to our system. As a case study we took Dublin City Councils project on intelligent Traffic System. First they tried with the basic big data software but they ran into many troubles so they joined IBM's research program. In this program they used IBM InfoSphere stream software which is used for the machine learning concepts. They used three different approaches to make the traffic system intelligent. They used instrumental approach which was collecting data from 1000 buses across the city and they are handling the data using the Info Sphere stream software. They used the digital maps and dashboards to keep track of all the buses at a glance. This is an interconnected approach. They used the system's intelligence to see the image congestion to identify the traffic congestion.
One approach to represent the system is as follows.

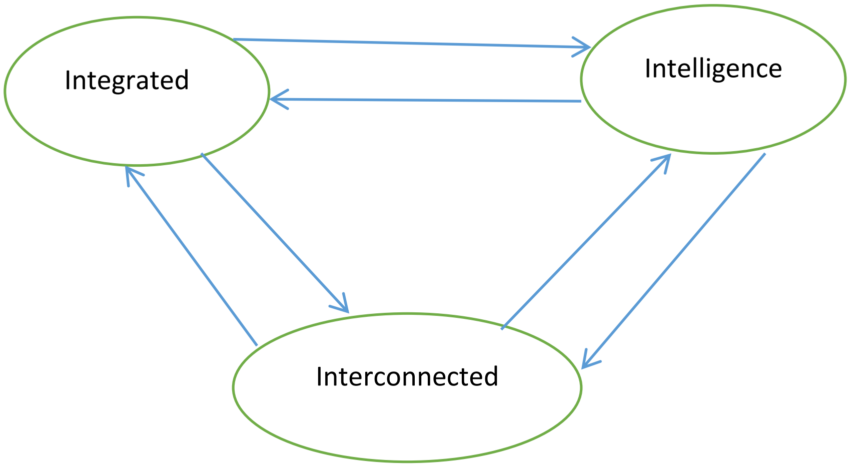


Figure 7.2.1

We are using this approach to make our system compatible in real time data using InfoSphere Streams.

7.2 South West Research Institute (SWRI)

South West Research Institute is a leader in the development, design and deployment in the Advance Traffic Management System. They are building software for different traffic related technologies.

Some of their systems are as below:

Advanced Traffic Management Subsystems

- Automated vehicle identification (AVI)
- Automated vehicle location (AVL)
- Incident detection and response plan generation
- License plate recognition
- Automated incident detection
- 911 incident integration
- 511 traffic conditions interface

SWRI is a big research institute which is located in Texas. They are doing research in many different fields. One of that research areas are Automation and Data Systems. They are developing an integrated System which includes Intelligent Transportation systems with different data gathering techniques to providing a real time solution to the user about the travel time, traffic congestion and traffic speed. From this case study we congregated the knowledge of the real time Transportation System which should have Traffic congestion indicator with travelling Speed and travel time facilities in it.

7.3 Zhejiang City Case Study

Trust Way is a company in China which is providing an intelligent traffic management system in the city of Zhejiang which is one of the developing cities. Due to city growth and industrial development, the city encountered traffic problems so government took a project to solve the problem with the help of TrustWay. This company is using Intel's Hardware support and using Apache Hadoop to solve the traffic problem.

They had different issues while developing the software like a centralized system which can gather data from different sources, handling the big data and optimizing the data for the use and guarding the traffic flows.

They solved these issues by combining the different technologies. They used Intel's Xeon processor E5 series which can handle the synchronous transmission of over 500 kb pictures with average speed of over 250 times per second. They used Apache Hadoop which enabled efficient techniques to analyze the data and finding the appropriate solution within second. For example one of the functionalities they are providing is to recognize the Number plates. With Apache Hadoop they can find the data within 2 second from more than 2.4 billion data.

Our focus to study this case was how we can manage big data with the hardware and how much our system should be capable of doing it.

Chapter 8. Conclusion

As we can observe, the traffic problem is a pretty generic problem encountered in almost every alternate city throughout the world. This problem can be efficiently managed by the use of latest technology by harnessing Big Data and Apache Hadoop at its full potential.

The city of San Jose is experiencing similar issues currently and we can definitely apply this model and with the help of authorities successfully implement the change in traffic management system making life easier for everybody in the city. Big Data Analytics make it possible for this solution to be successful in the long run owing to its flexibility, adaptability and intelligence. Thus, enabling room for future customizations in lieu of changing requirements.

With the help of the customizable and powerful tooling provided by IBM infoSphere Streams we learnt that stream computing can be explored as a viable option for performing real time data analytics. It's easy-to-use and wide range of processing capabilities into meaningful datasets help in data management as redundant data is filtered at the initial stages itself.

Thus, we can positively conclude from that this system can used to solve the traffic problems in currently developing cities as well as serve as a guideline to the consumer. Our system can provide a real time traffic solution by integrating real time data and the predictions generated from the past data stored. The system will able to provide a user friendly environment and an easy solution for the traffic jams, traffic speed, traffic flows by diverting the traffic to appropriate routes. There are various other features that can be integrated with the application like emergency alerts, weather data, and digital data to conclude the predictions and this will be helpful for the user to know various details as an all in one facility.

Chapter 9. Bibliography

Wikipedia-Big Data

URL : http://en.wikipedia.org/wiki/Big_data

How Stuff Works-Big Data

URL:http://computer.howstuffworks.com/internet/basics/what-is-big-data-.htm

Big Data by O'Reilly

URL: http://radar.oreilly.com/2012/01/what-is-big-data.html

O'ReillyMedia.

URL:http://www.forbes.com/sites/oreillymedia/2012/01/19/volume-velocity-variety-what-you-need-to-know-about-big-data/2/

IBM Big Data Analytics.
URL: http://www.ibm.com/analytics/us/en/

IBM InfoSphere Streams

URL:www.**ibm**.com/software/products/en/**infosphere-streams**

Stream Computing

URL:www.**ibm**.com/software/data/**infosphere**/**stream**-computing/

Forbes.

URL: http://www.forbes.com/sites/ciocentral/2012/04/16/the-big-cost-of-big-data/

Apache Hadoop.

URL: http://hadoop.apache.org/

South West Research Case Study

URL:www.**swri**.org/3pubs/brochure/d10/ITS/its.pdf

Dublin City Case Study

URL:http://www.ibm.com/midmarket/ie/en/att/pdf/IBM_DCC_130715.pdf

Zhejiang City Case Study

URL: http://www.intel.com/content/dam/www/public/us/en/documents/case-studies/big-data-xeon-e5-trustway-case-study.pdf