Understanding Big Data through Visualization and Geospatial Analysis

A Technical Paper prepared for the Society of Cable Telecommunications Engineers

By

Randall Frantz
Director – Telecommunications Solutions
Esri
380 New York Street
Redlands, CA 92373

+1 (909) 793-2852x1-2958
rfrantz@esri.com

Overview

It is said that we live in the information age. However, in reality we are more likely to drown in data, specifically Big Data, than sail smoothly on a sea of information. Big data, as defined by Gartner, is when the three “V’s” (Volume, Velocity and/or Variety) overwhelm the ability to process and transform data into the actionable intelligence required to support business decision making processes. Data must be given structure and provided context for it to be useful and actionable. Today’s technology provides a means to collect every social media exchange, financial transaction or customer activity on a daily basis creating vast amounts of data. To effectively leverage all of this data and gain a competitive advantage in the marketplace, new tools are required.

This paper will explain how the visualization and spatial analysis capabilities of a Geographic Information System (GIS) can cut big data down to size. We will explore how the inherent capabilities of GIS technology, used to manage, organize and process vast amounts of spatial data, can be applied to much of the data that is generated by our Operational Support Systems (OSS) and Business Support Systems (BSS).

Providing a spatial context to this data can deliver an unparalleled level of granularity and uncovers new relationships and correlations between activities that often can only be revealed through spatial analysis.

Discussing these capabilities will provide a greater understanding of how GIS technology can turn the challenge of big data management into beneficial and actionable intelligence.

Big Data Characteristics

It is estimated that the world generated over 2.8 zettabytes (2.8 x 10^{21}) of information in 2012. With the annual growth rate of 40-50%, this could reach 40 ZB by 2020. To put this statistic into perspective, storing a zettabyte of information requires 250 Billion DVD's. This equates to over 98 DVD's of data for every person on earth in 2012 and an equivalent 1,250 DVD's per capita by 2020.

Where will this data come from? In the past data was often collected manually which required physically inputting data into a storage system. The emergence of the digital age reduced the cost of data acquisition, since most data can be collected through remote sensing and other electronic means. Every transaction on the internet and every phone call generates data. Satellites circle the globe constantly collecting imagery, be it weather or a variety of other types of data. With the cost of data collection declining, the mere thought that it might be useful at some future date often justifies the collection and storage of it. Think of the black box in an airplane. Most of the data is never used but it is collected so that, in the rare instance of an accident, the data will be available to analyze. There is even talk about putting black boxes in new cars, something that would have been cost prohibitive in the past.

As previously mentioned, Gartner uses 3Vs to describe big data: Volume, Variety and Velocity. When these 3Vs reach a high level they require special processing. The proliferation of electronic and remote sensing devices provides an exponential increase in the volume of data collected. It also results in a wide variety of data. The black box in an aircraft provides an example of the 3Vs. It rapidly records a large number and a wide variety of data points. There are hundreds of sensors throughout the plane. The engine sensors record data on intake and exhaust temperature, throttle setting, and thrust, to name just a few. The black box also collects data from other sensors that record pressure, airspeed, control surface deflection, altitude, fuel levels, etc. All these data points are important in understanding how the aircraft is operating. The data must be collected rapidly with a very short interval between sample sets. Sampling this aircraft data at 1 hour intervals is not of much value in a post incident reconstruction. However, when collected in microsecond intervals it can provide a wealth of information. Each individual observation is worth a little less since the changes will be smaller but the set of all observations increases in value in relation to the data collected. By analyzing all these data points it is possible to reconstruct, in great detail, the events leading up to an incident, provided you have a model and a system that can accurately process the data.

Companies collect vast amounts of data with the expectation that managers who better understand organizational performance will ultimately make better business decisions. Dynamic businesses, such as telecommunications, require accurate and timely data in order to respond quickly in highly competitive markets. In the past, managers often received incomplete and outdated data. The thinking has been if a little data improves the decision making process, then more data must lead to even better decisions.

Today the opposite is true. Managers now have access to large volumes of data. The problem with collecting so much data is that it becomes impossible to process and turn it into useful information and actionable intelligence. When managers become overwhelmed with data, their ability to make better decisions is only slightly better than when they had too little data. If the data could be processed, organized and analyzed, it would provide a wealth of information on how the business is performing and significantly improve decision making.

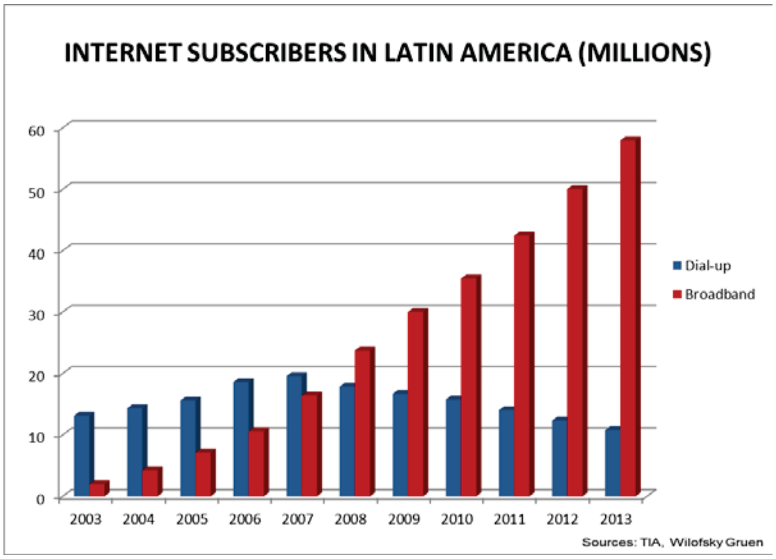
Before data can be useful, it must be transformed into information. Information is created when data is organized and structured in a way that creates context. Visualization, in the form of graphs and charts, is an effective method of improving our understanding of data. It works well for simple analysis, but users can become overwhelmed when the volume and complexity of graphs and charts grows to the point that it assumes the characteristics of big data and becomes too complex, overwhelming and unmanageable.

Power of Geographic Information System

Most data collected has a spatial component and telecommunications is a business that is heavily location dependent. The service type, quality of service and customer experience is dependent on where you live or work along with the network capabilities serving you. For example, with wireless service the quality and bandwidth available might depend on whether you are located inside or outside a building. By including location into the analysis, a wireless provider can understand that a high drop call rate was due to the lack of coverage inside the building. This understanding provides the actionable intelligence necessary to effectively address the problem. Adding a new outside cell tower or increasing its capacity will not solve the problem. Solving the service problem might require a picocell inside the building. Just as in the use of graphs and charts, adding visualization with geospatial analytics can help solve this problem by significantly improving informational context.

An excellent tool for visualization and spatial analytics is a GIS (Geographic Information System). A fully functional GIS contains very powerful tools designed specifically to process and analyze spatial data. The power that GIS delivers is the ability to both visualize and analyze data spatially. Many BI tools provide data visualization in the form of graphs and charts. Visualization of the data greatly increases our ability to see trends. It also allows us to more quickly comprehend not only changes in the data but to quickly access other factors such as the rate of change (the slope) and the acceleration in the rate of change (straight line versus curved slope).

Figure 1 is a chart that illustrates the growth of broadband and decline of dial-up subscribers in Latin America. Sifting through the tabular data would have revealed that broadband grew during the study period and dial-up declined. However, the chart reveals information about the growth and trends of internet subscribers more quickly. The chart reveals that dial-up was growing until 2008 and then began to decline while broadband is growing at an accelerated rate.


INTERNET SUBSCRIBERS IN LATIN AMERICA (MILLIONS)

Sources: TIA, Wilofsky Gruen

Figure 1

Other visualization capabilities of GIS, specifically spatial tools, can provide more context to the data. Latin America is a very large geography. A color coded or thematic map could provide country level information illustrating where broadband is growing the fastest or dial-up declining more rapidly. A GIS can even perform temporal analysis which shows the change of internet access technology over a period of time by country.

Companies such as Softbank in Japan have used temporal analysis to create traffic maps which show the concentration and dynamic flow of mobile phone users over a typical business day. This type of analysis clearly shows the migration from the suburbs into the city and back to the suburbs during a 24 hour period. It is possible to see the concentration of users as they travel from their homes to transportation hubs, commute into the city and reverse the pattern in the evenings. By studying this daily migration pattern engineers can predict and plan network capacity requirements. By using forecasts of expected population growth and demographic information they can also predict where and when additional capacity will be needed. Softbank uses GIS technology to analyze customer requirements and compare them against their existing network technology and capacity to help plan their capital investment.

To fully appreciate how a GIS can process data requires understanding the spatial capabilities of a GIS and its ability to integrate many types of data using the common element of geography. A GIS has the ability to draw data from multiple sources and create relationships between layers of data based on location.

Figure 2 illustrates multiple layers of data including imagery, features, network infrastructure, service areas, customers and demand. Typically, this data is stored in many different systems within a company often making it difficult to model the relationship between the various layers of data. The data from these different systems can be extracted and combined and organized into layers. Once the layers are created, the GIS can analyze the data relationships between them.

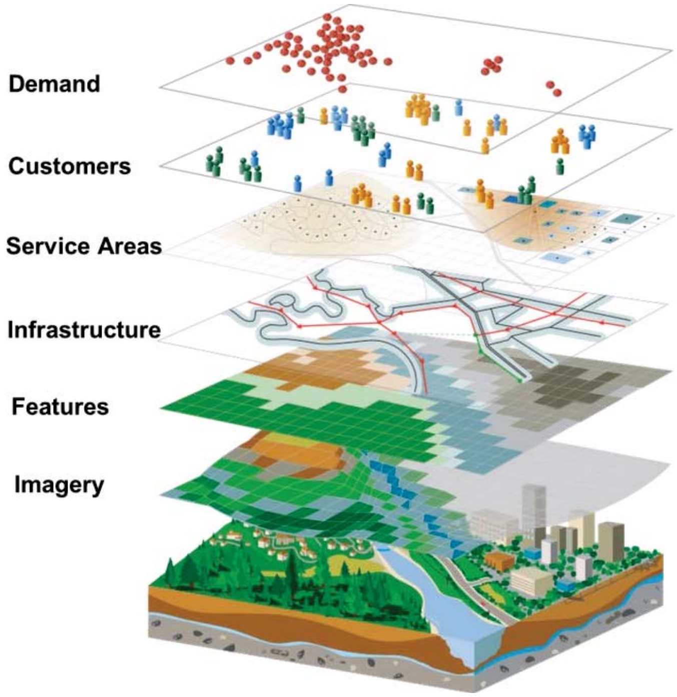
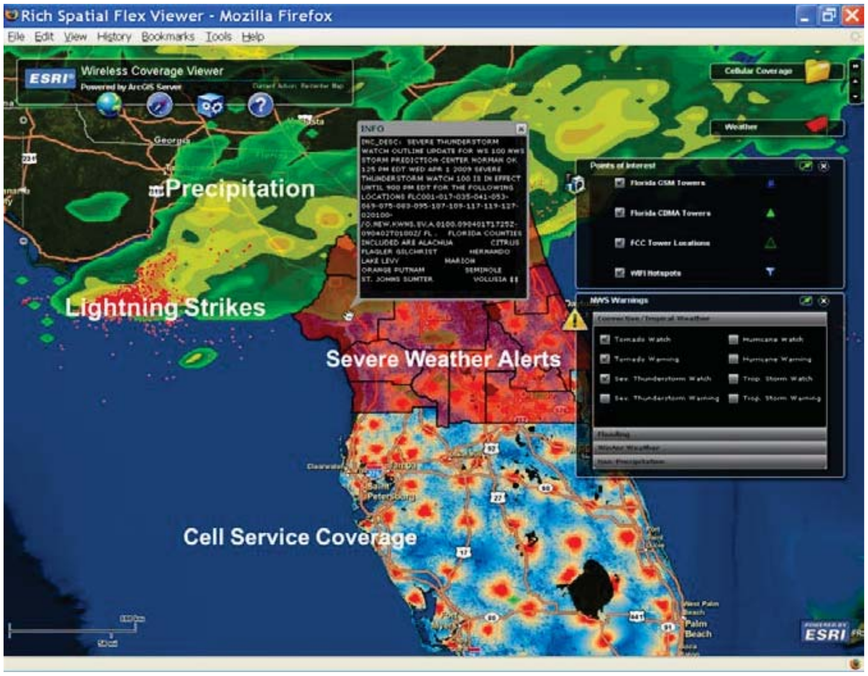


Figure 2 – GIS Data layers

A GIS can handle a wide variety of spatial data using location as the common thread.

Figure 3 shows how various layers of data from multiple sources can be combined into a single view. The geographic region shown is Florida with several data overlays from different sources including precipitation, lightning strikes, severe weather warning alerts and raw RF cell coverage. The ability to view and understand the wealth of data provided through this single view provides operators with information needed to understand threats to their network and manage their assets accordingly.



INFO

INC_DESC: SEVERE THUNDERSTORM WATCH OUTLINE UPDATE FOR WS 100 NWS STORM PREDICTION CENTER NORMAN OK 125 PM EDT WED APR 1 2009 SEVERE THUNDERSTORM WATCH 100 IS IN EFFECT UNTIL 900 PM EDT FOR THE FOLLOWING LOCATIONS FLC001-017-035-041-053-069-075-083-095-107-109-117-119-127-020100-
/O.NEW.KWNS.SV.A.0100.090401T1725Z-090402T0100Z/ FL . FLORIDA COUNTIES INCLUDED ARE ALACHUA CITRUS
FLAGLER GILCHRIST HERNANDO
LAKE LEVY MARION
ORANGE PUTNAM SEMINOLE
ST. JOHNS SUMTER VOLUSIA $$

Points of Interest
☑ Florida GSM Towers
☑ Florida CDMA Towers
☑ FCC Tower Locations
☑ WiFi Hotspots

NWS Warnings
Convection / Tropical Weather
☑ Tornado Watch
☑ Tornado Warning
☑ Sev. Thunderstorm Watch
□ Sev. Thunderstorm Warning
□ Hurricane Watch
□ Hurricane Warning
□ Trop. Storm Watch
□ Trop. Storm Warning
Flooding
Winter Weather
Non Precipitation

Figure 3 – Integrating Data Layers from Multiple Sources

Data and Spatial Tools

There are two basic data types in a GIS, vector and raster data. Vector data represents geographic features such as points, lines and polygons. Attributes are associated with each vector feature. Vector data can be used to represent poles, fiber nodes, cables, service territories, etc. Raster data is an array of equally sized cells arranged in rows and columns. A typical example of raster data is satellite imagery. The images are composed of individual tiles that create the full image. If you zoom in too close, the images become tiled. The GIS can perform geoprocessing functions applied against vector and raster data.

Cellular companies use geoprocessing tools against both vector and raster data to create simplified and understandable service coverage maps from a very complex Radio Frequency (RF) propagation analysis. Cellular companies use various types of RF design and planning tools to calculate the coverage of a single cell tower. The raw engineering data for a single tower is represented in the Figure 4. Signal strength which is typically raster data ranges in value (-100 to 56) and is represented by color with red being the strongest and blue the weakest. Signal strength is calculated based on type of antenna, frequency, power output and terrestrial features that might obstruct the signal.

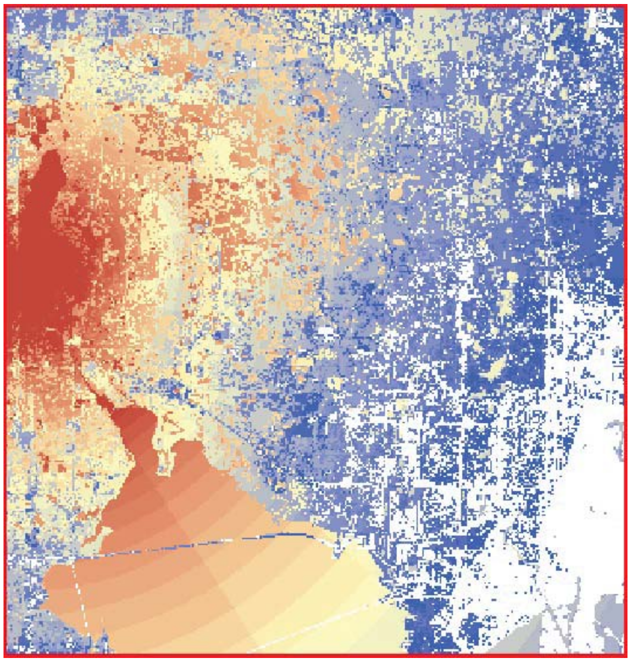


Figure 4 – Raw RF Signal Strength

Coverage maps are typically used by internal service provider sales personnel or can be public facing to answer customer inquiries about general service coverage areas. The detail raw RF signal coverage is too detailed and complex to be used by these groups as a general coverage map. A GIS is required to transform this complex dataset into a simple coverage map that can be used by internal sales personnel or the general public. The spatial tools available can help achieve the following goals:

- Eliminate coverage over areas of water
- Remove outlier pixels (areas of meaningless or insignificant coverage)
- Smooth output to remove pixelation
- Create fixed bands representing quality of service

For example, the first step is to remove the coverage predictions over water which is typically not included in coverage maps. The next step is to use a polygon to raster tool to convert the water polygons to a raster data which will be used as a mask to reset the signal strength values over the water to NoData, see Figure 5. This will remove the coverage over water.

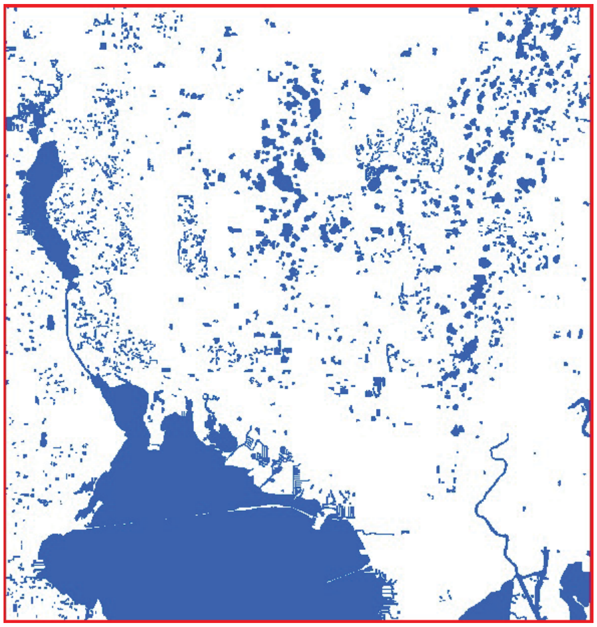


Figure 5 – Water Coverage Mask

With the water signal strength values removed the next step is to reduce the details by using the Reclassify tool. This tool will generalize the data and remove much of the detail. The reduction in data will simplify the ability to process using the remaining tools. We have chosen to reclassify to five levels which is the goal we outlined in the process. The next step is to apply the Majority Filter tool which will replace raster cells based on the majority of contiguous neighboring cells. This will have a smoothing effect. See Figure 6 for the results of performing both functions.

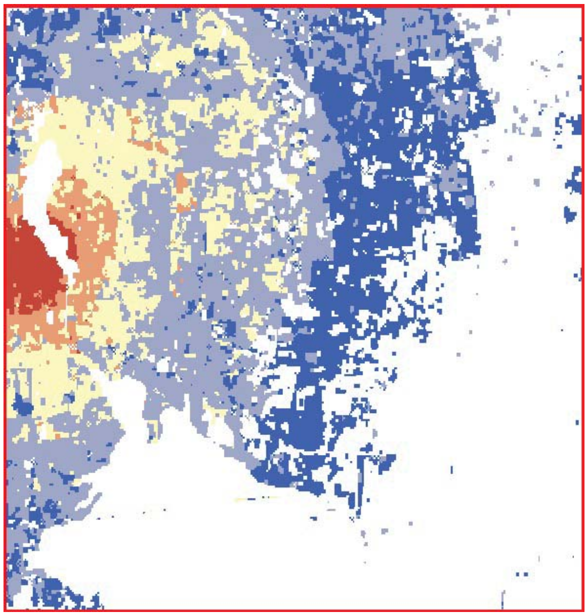


Figure 6 – Reclassify and Majority Filter

The final step in the process is to convert from raster to vector, creating polygon features that provide the general service bands desired. These bands can be color coded, see Figure 7, from blue to red indicating the signal (poor, fair, good, very good and excellent) identified in our goals. A more comprehensive technical discussion of the RF Cleanup process upon which this section was based was developed by Mark Reddick, Telecommunications Technical Marketing Lead at Esri and can be found at http://blogs.esri.com/esri/arcgis/2010/10/12/rf-propagation-cleanup-tools/.

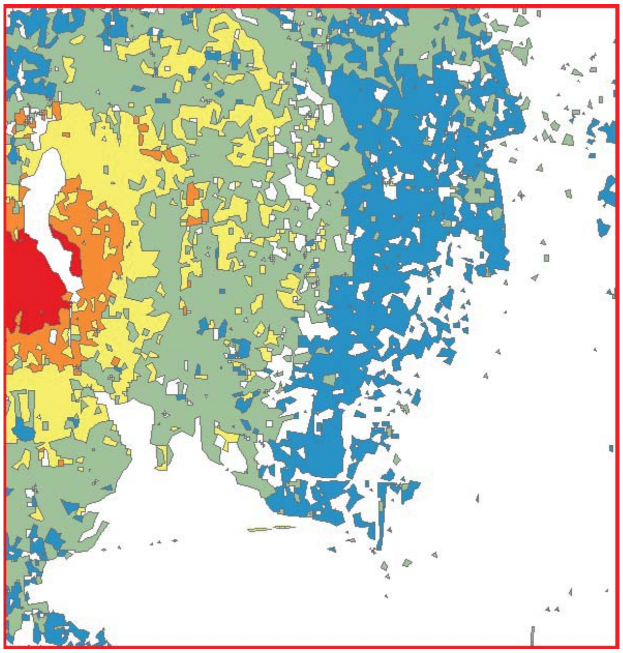


Figure 7 – Color Coded Service Strength

This process provides the banded coverage for a single tower but the process can be run on larger datasets including signal strength output for an entire service area. If the resulting coverage map is combined with a geocoder, a user can enter an address which will be placed in the appropriate polygon which will then identify one of the designated service levels. The same analysis can be applied to HFC, COAX or fiber networks. Many of the broadband coverage maps deployed in the United States during the past several years use the point and polygon analysis to identify broadband service availability.

Hadoop and MapReduce

As data volumes have grown, individual computing systems and databases have not scaled equivalently. The single computer accessing and processing on single-node databases cannot handle big data. Big data has forced an evolution to distributed systems where data is stored across a number of servers or file systems (Figure 8). Processing takes place on the server where the data (or a portion of the data) is stored.

Google was a leader in this field implementing their enterprise on a distributed network, composed of many nodes. Google then fused storage and processing into each node.

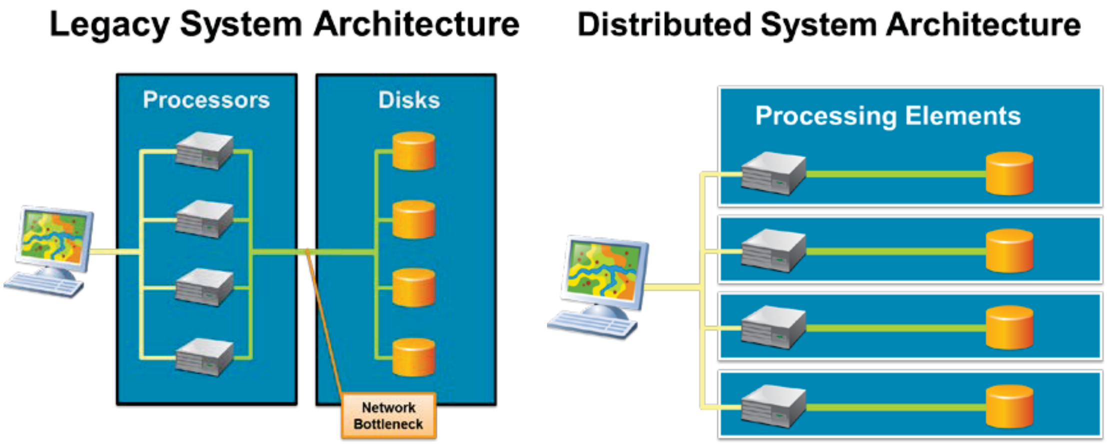


Figure 8 – Legacy vs. Distributed System Architecture

Hadoop is an open source data processing system implementation of the Google framework. It supports a distributed system for both storage and computation of data. GIS systems are beginning to make use of this same paradigm so that they can work with very large datasets. The ability to leverage Hadoop's capabilities significantly improves response times. Hadoop can run processing tasks on large hardware clusters. Hadoop uses MapReduce to divide these tasks into many smaller tasks. These tasks can then be executed on any node within the cluster. The MapReduce system manages the data transfers and communications so the tasks can be run in parallel. This system provides the ability to execute operations that require massive amounts of data and computation.

An instance of a MapReduce program is called a job. A high level MapReduce walk-through is shown in Figure 9. The job accepts arguments for data input and outputs. The MapReduce framework splits the data into map() functions that are run in parallel. The framework runs a Combine to join map() outputs and then performs a shuffle and sort on the data. The reduce() functions work against the sorted data and writes a part of the results to a file.

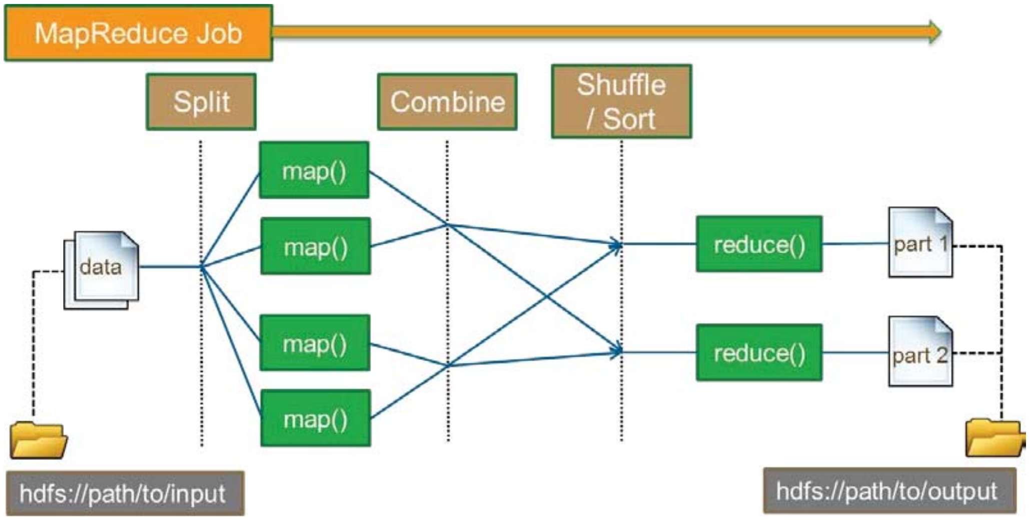


Figure 9 – High Level MapReduce Walk-Through

The ability of a GIS to leverage the cooperative processing available in MapReduce significantly increases the ability to handle extremely large datasets. An example of using MapReduce for a polygon count is provided in Figure 10. The application is designed to count the number of polygons represented by three states: Washington, Alaska and Oregon. The example is simplistic but with larger datasets and more complex analysis, the savings can be significant. Spatial analysis using this framework can reduce files with 3.5 billion records to 1.1 million and improve response time from hours to minutes.

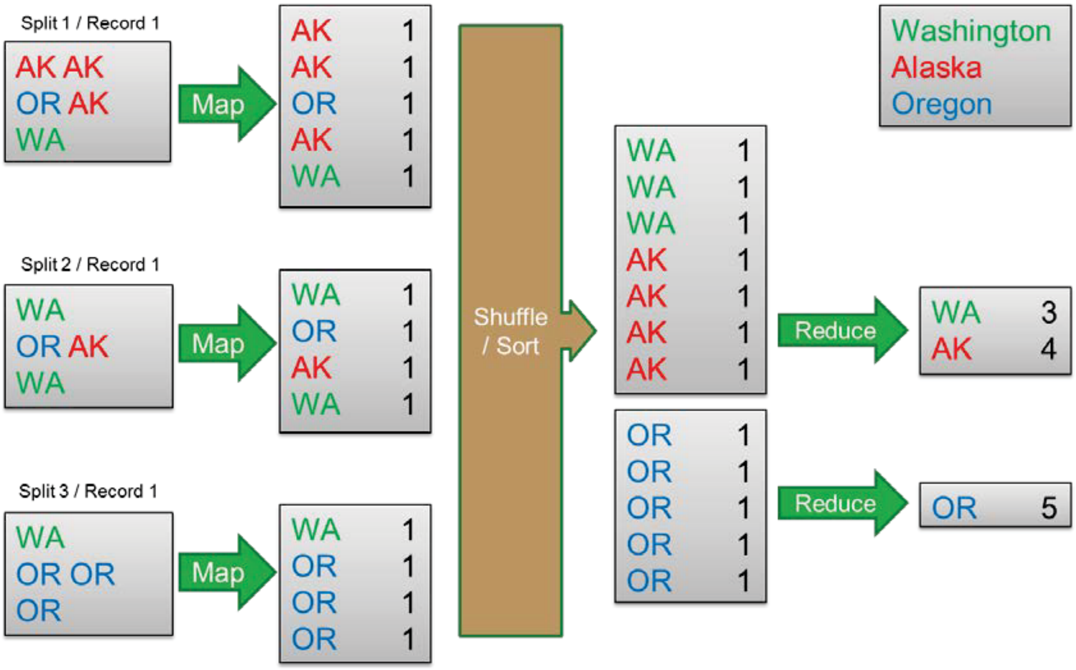


Figure 10 – MapReduce Polygon Count

Real-time GIS

GIS is a platform for understanding real world conditions. When monitoring the health of the network or responding to threats such as natural disasters, it is essential to transform the data collected into real-time information. As mentioned earlier, today's remote sensors and network operational support systems generate high volumes of data. Even social media can provide rich and useful data sources. If data can be quickly transformed into actionable intelligence, managers will have the capability to make more timely and accurate decisions. In the telecommunications industry this can save millions of dollars. In other industries such as law enforcement and firefighting it can save lives.

Real-time GIS is an essential component for responding to natural disasters or other emergencies. GIS can work with data from many authoritative sources and integrate this data using location as a common factor. Real-time GIS data is a continuous stream of events flowing from sensors where each event represents the latest state of a sensor used to create Features as displayed in Figure 11. The applications determine what data should be collected and processed as well as the time intervals of collection.

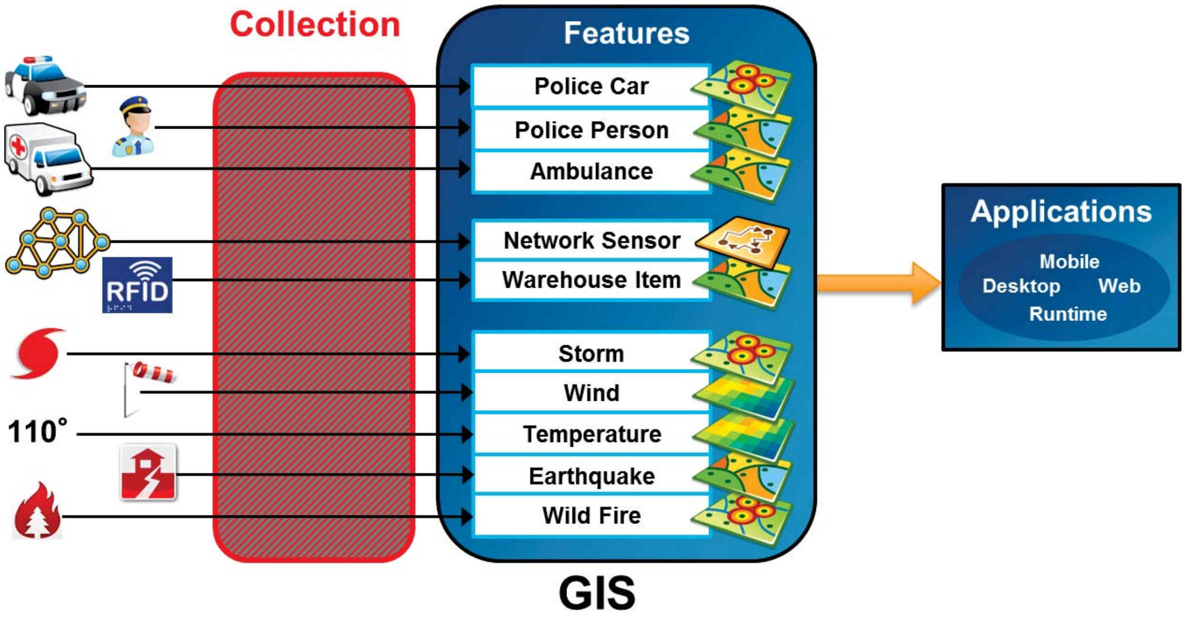


Figure 11 – Real-time GIS Data

Fighting wild fires, which are common in the southwestern United States, provides an excellent example of using a GIS platform to process and analyze real-time spatial data.

Information is collected from multiple sources. Satellites provide thermal imagery to identify hot spots. They along with other sensors provide information on weather forecasts, including wind speed, temperature and humidity. This data can be overlaid on terrain and vegetation data to forecast the fires path and burn rate. Combining this information with the availability and location of firefighting resources provides the wildfire command center with the information they need for timely evacuation and deployment of resources.

The same system can be used to understand network performance and the impact of network failures on service levels as illustrated in Figure 12. The network sensors and alarms are fed into the GIS which maps the status of the network elements. When the status of the network changes the map is updated, depending on the predefined triggers. Those events labeled 'noncritical' can be stored and corrected during the next scheduled maintenance. However, those events which are 'major' will trigger the launch of notifications to the appropriate personnel.

As illustrated in Figure 12, when a 'major' network outage such as a node failure is detected additional processes can be launched automatically. In this example, the coverage map is redrawn highlighting the affected area. Additional information is extracted from other data sources to display not only the affected areas but the total population in the affected area. This is important information when allocating limited resources and developing prioritization criteria for service restoration.

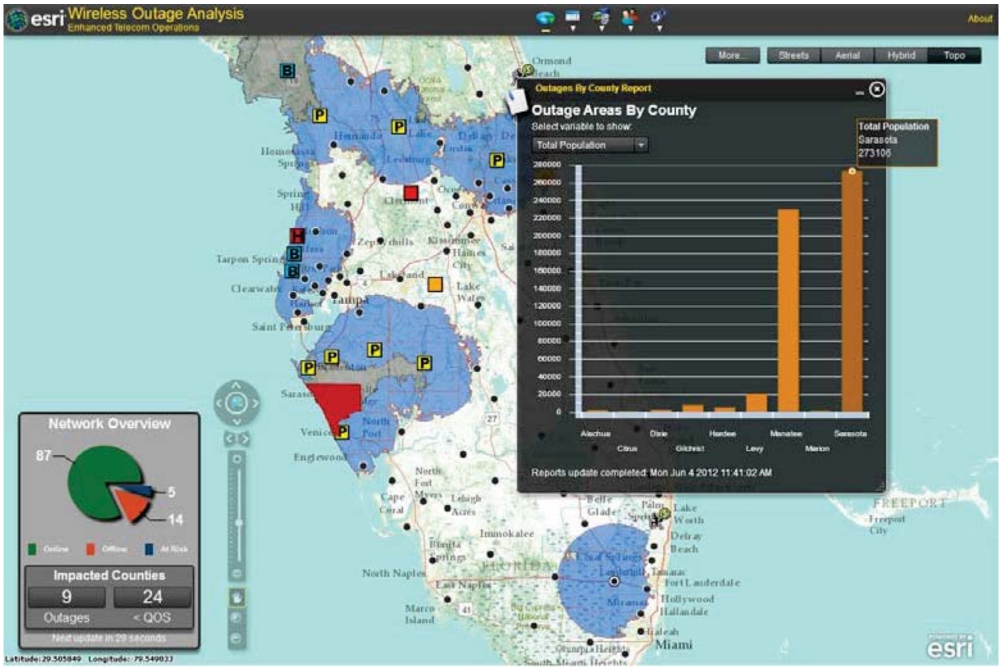


Figure 12 – Outage Coverage Map

The processing and management of real-time data by the GIS can be extended to feed other systems as displayed in Figure 13. Connectors can be created to feed downstream systems such as another GIS used for additional geoprocessing or external platforms like Twitter or Hadoop. The process can be repeated as often as necessary to meet specific business needs.

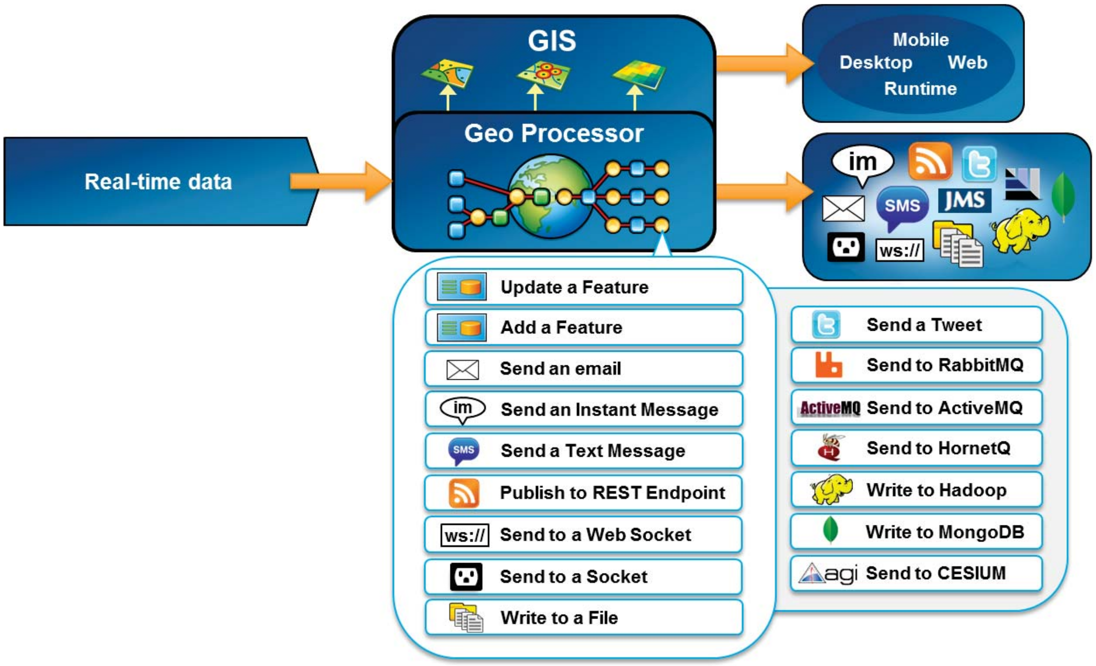


Figure 13 – Extending Real-Time Data

Summary

GIS is increasingly becoming a powerful platform used by many industries to process, analyze and derive value from big data. It is a resource that can help transform big data into actionable intelligence. GIS has and continues to evolve into a timely and useful solution that leverages the unique capabilities of location to uncover hidden relationships within data. And, it provides tools so managers can make timely and accurate business decisions based on a comprehensive understanding of organizational performance.

A GIS can address the 3Vs of big data: Volume, Velocity and/or Variety. Large volumes of data can be handled by leveraging the Hadoop Distributed File System and MapReduce platform. The geoprocessing tools in a GIS can be used to focus on specific geographic areas of interest regardless of political or geopolitical boundaries. The variety aspect of big data is managed by creating layers of information where the common thread is location. GIS enables these layers to be integrated and analyzed. Finally, the velocity of today's data is being processed with real-time GIS tools.

As big data grows, the capabilities of GIS technology will evolve, scale and grow as well, ensuring that tomorrow's big data challenges can be transformed into recognizable, manageable and actionable business opportunities.