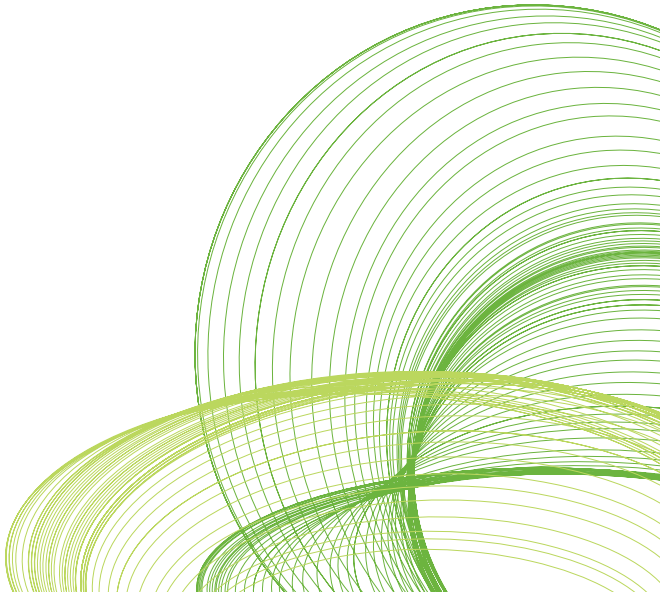
QlikView

QLIKVIEW AND BIG DATA

A QlikView Technology White Paper

July 2012

qlikview.com



Introduction

There is an incredible amount of interest in the topic of 'Big Data' at present: for many organizations its use is an operational reality, providing unprecedented ability to store and analyze incredible volumes of disparate data that are critical to the organization's competitive success, enabling people to identify new opportunities and solve problems they haven't been able to solve before. For many others, Big Data is a big trend in present-day IT that needs understanding and its relevance needs to be separated from the hype surrounding the topic. This paper discusses the role of the QlikView Business Discovery platform as the foremost analytics platform accompanying a Big Data solution. It is written for both IT professionals and business leaders who are trying to understand how to gain the most leverage from a Big Data implementation by providing an analytics layer that can both access the data and make it relevant and accessible to the business users in an organization.

Why Big Data Matters

In the U.S. alone there are roughly 9 million airplane flights a year, each of which generates data about hundreds of parameters every second or so, from the aircraft and radars and other sources. In addition, each flight has unstructured data associated with it such as safety reports and reports from pilots and co-pilots.[i] NASA (the National Aeronautics and Space Administration) is using analytics to dig through all this data and gather insights that can identify potential runway incursions and other accidents and prevent them before they occur.

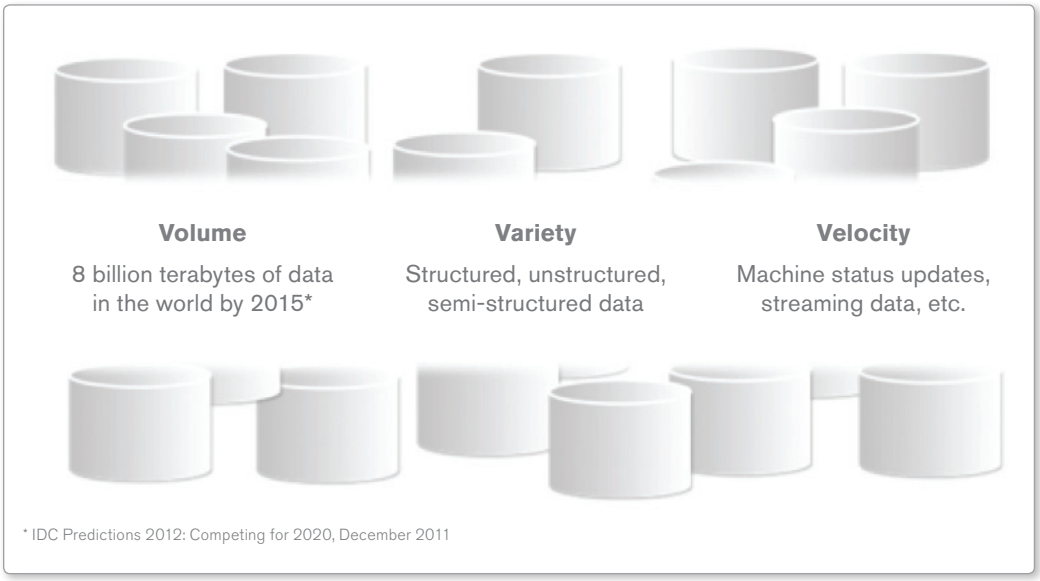
In another example, eCommerce giant eBay has a social data intelligence program in place to help decision makers better understand the company's audiences, influencers, and competitive position, and deliver superior customer service. As of June 2012, eBay had indexed more than 40 million blogs and forums (60 billion posts – 10,000 a second!), which amounts to 65 terabytes of data. A global social analytics team works with multiple groups across the company to find and share insights from all this data.[ii]

For commercial entities like eBay, the more data the organization can manage and analyze compared to its competition, the greater its competitive opportunity. For public organizations like NASA, the more data it can process and analyze, the more accurate its predictions can be. These are just a couple of examples of why Big Data matters.

What is “Big Data?”

According to McKinsey Global Institute and others, the term “big data” refers to data sets whose size is beyond the ability of typical database software tools to capture, manage, and process within a tolerable elapsed time. Depending on the industry, this can mean data sets ranging from a few dozen terabytes to multiple petabytes. In addition, the term Big Data is associated not only with the volume of data but also the variety (e.g., the types of data, structured or unstructured, etc.) and velocity (the dynamic or changing nature of the data as new data flows into, and old data exits, a system). To put it simply: if it’s too big to manage, it’s Big Data (see Figure 1).

FIgure 1. What is Big Data?

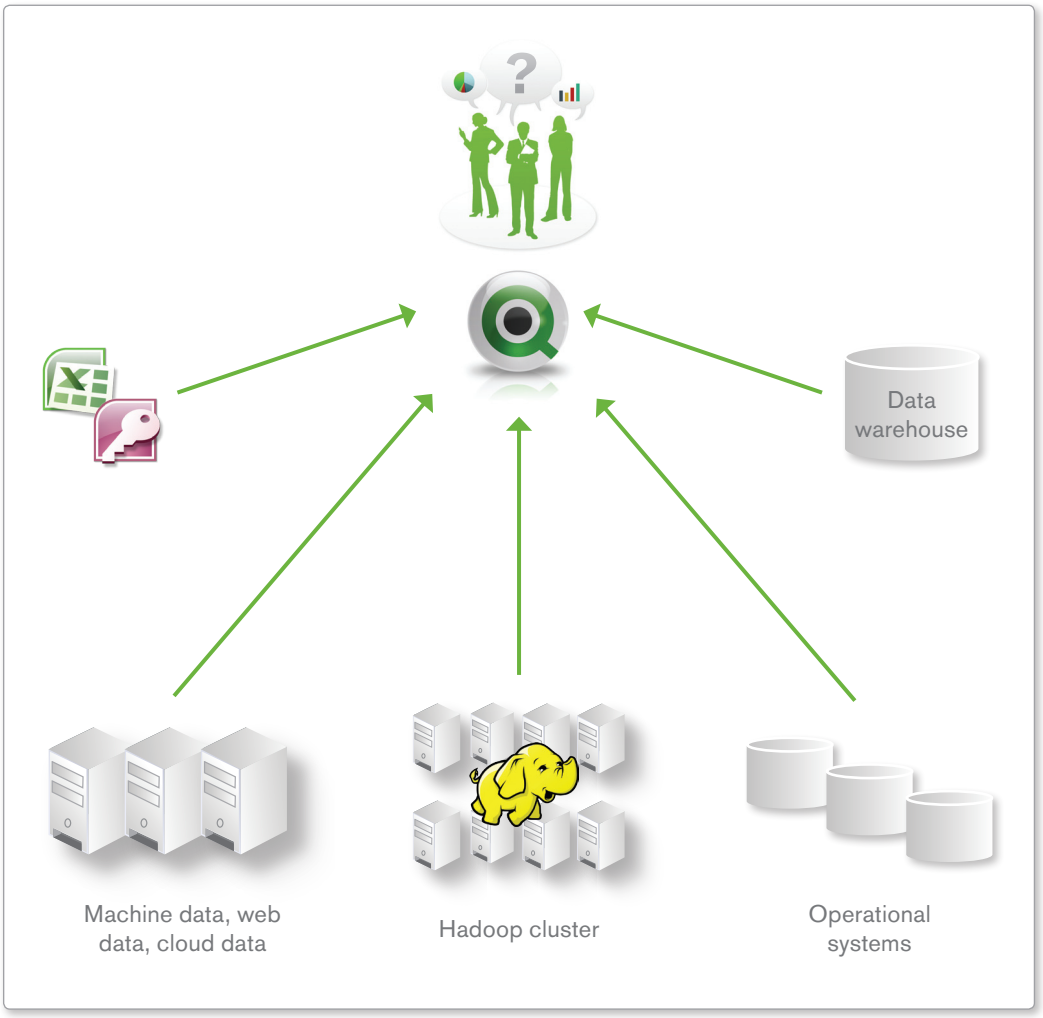


In recent years the world has seen the rise of Big Data solutions such as Hadoop and Google BigQuery. What these have in common is they utilize distributed computing networks or massively parallel processors (MPP) to provide storage of and access to extremely large amounts of data. This paper is not intended to discuss the underlying technologies associated with Big Data providers, although it will offer a brief technical overview in a later section.

One of the important characteristics of Big Data is that it is often used to store and process unstructured data (e.g., web content such as online reviews, text, social media content) in addition to structured (but highly voluminous) machine data such as data sourced from sensors (e.g., electricity meters), or automated computer systems (e.g., computer logs or algorithmic stock transaction systems). In order to make practical use of such data, organizations would like to be able to marry this data with existing structured data from their internal OLTP (online transaction processing) systems, data warehouses, and enterprise systems like CRM (customer relationship management) and ERP (enterprise resource planning). By bringing this data together and being able to identify patterns and associations, they can conduct analysis into customer sentiment, customer behavioral patterns, product quality or safety issues, and clinical trial effectiveness.

QlikView plays a critical role in Big Data implementations, providing both the rapid, flexible analytics on the front end as well as the ability to integrate data from multiple sources (e.g., the Big Data source, existing data warehouse, departmental databases, and spreadsheets) in one single, interactive analytics layer (see Figure 2).

Figure 2: Insight Comes from Big Data in Context with Other Data



Big Data – Complimenting Existing Information Architectures

Big Data systems are most commonly seen as complimentary to an organization's existing data infrastructure rather than as a replacement. Some of the reasons for this include:

- Significant historic investment in centralized and localized data warehouses
- Concerns about data security and protection—unwillingness to put sensitive or valuable data in a cloud-based environment
- Complexity of using existing software tools to deploy and manage Big Data systems like Hadoop; the current generation of Big Data management tools is less evolved than standard SQL data access tools
- Batch-based processing of queries rather than dynamic queries means high latency (poor performance) that is unacceptable for business-user analytic requirements
- Shortage of IT skills with Big Data management systems like Hadoop, NoSQL, and Google BigQuery
- Big Data management systems like Hadoop, NoSQL, and Google BigQuery would be overkill for many data challenges; standard RDBMSs are sufficient.
- Big Data management technologies are new and many are open source and/or in beta.

The Role of Analytics in Big Data

Big Data solutions in isolation provide little value to an organization unless that data can be acted upon to support the decision-making process. While much has been written and talked about the underlying technologies that offer storage of and access to extreme volumes and variations of data using distributed computing capabilities, it is only in the analysis of that data that real value is extracted. While this is true for data of any size or type, it is particularly relevant in the Big Data arena. As Wayne Eckerson from TechTarget points out in his research paper entitled "Big Data Analytics: Profiling the Use of Analytical Platforms in User Organizations", "A valuable characteristic of 'big' data is that it contains more patters and interesting anomalies than 'small' data. Thus, organizations can gain greater value by mining larger data volumes than small ones."

Data Staging for Analytics: Making It Relevant

One important characteristic of Big Data solutions is that, almost by definition, it is inevitable that much of the data contained in the system is useless, mere "pass-through" data accumulated simply because it can be. As a result, attempting to provide an analytics layer to accommodate all of the data is in many cases an unnecessary approach. In many cases, it has been shown to be more useful to provide an aggregation mechanism to extract the most relevant and useful data from the Big Data source in preparation for analysis.[iii]

The Big Data solution should be considered as one part of the organization's overall information management architecture, working alongside existing data warehouses, Complex Event Processing (CEP) engines, analytical sandboxes, OLTP systems and so forth. As Eckerson mentions in his paper, "Today some companies use Hadoop as a staging area for unstructured and semi-structured data before loading it into a data warehouse." The Big Data system maintains all of the detail-level data while lightly-summarized data sets are made available to the analytics layer.

In this framework, QlikView fits naturally as either a direct recipient of the data from Hadoop or other Big Data systems, or running on top of the data warehouse, or both. In either case, QlikView's simplicity in terms of joining data from multiple sources for the purpose of associative analytics is very much evident.

QlikView and Big Data

Business users are constantly being challenged to efficiently access, filter, and analyze data — and gain insight from it — without using data analytics solutions that require specialized skills. They need better, easier ways to navigate through the massive amounts of data to find what's relevant to them, and to get answers to their specific business questions so they can make better business decisions more quickly.

The growth in adoption of massively parallel processing solutions for handling ever larger volumes of data — whether structured or unstructured — is driving demand for analysis tools to enable business users to derive insights from Big Data.

QlikView takes a two-pronged approach to this challenge.

First, QlikView's approach has always been to understand what it is that business users require from their analysis, rather than to force-feed a solution that might not be appropriate. Providing appropriate data for the use case is more valuable to users than providing all the data, all the time. For example, local bank branch managers may want to understand the sales, customer intelligence, and market dynamics in their branch catchment area, rather than for the entire nationwide branch network. With a simple consideration like this, the conversation moves from one of large data to one of relevance and value.

In any organization, the number of people who need to analyze extremely large data volumes is typically relatively small. For example, a retail bank might have thousands of branches; however, there may be only a hundred business analysts in a centralized, corporate role. While branch managers only need slices of data that are relevant to their operations, the corporate analysts may need access to much larger data volumes.

QlikView is designed to accommodate both types of use cases and enables users to focus on the data that is relevant to them and is of the highest value to them and their area of interest. By taking appropriate slices of the data – big or small – QlikView serves as an analytical app platform downstream of the data sources, to provide business analysts and less technical business users alike the insight they need from the data that is most relevant to them.

Second, QlikView has been addressing, and continues to address, the Big Data challenge by ensuring that targeted QlikView apps can address the amounts of data that are needed to ensure the relevancy of the app for business users. Here's how:

- Recent trends in large memory available on standard Intel hardware allow QlikView to handle ever-larger volumes of data in memory (which provides users with a super-fast, interactive experience).
- QlikView best practices promote an architecture-led deployment when handing very large data volumes, such as making proper use of distributed servers in a clustered environment; constructing appropriate apps for the intended audience; using sophisticated data reload engines; and using document chaining where necessary to allow aggregated views to be coupled with detail-level views while optimizing hardware resources.
- QlikView provides an open data protocol (QVX, or QlikView data exchange) via a series of APIs (application programming interfaces) developers use to interface with the APIs of Hadoop- and other Big Data system providers. QlikView's QVX protocol can be used to connect to Hadoop based systems via two different methods:
    - Disk-based QVX file extracts from Hadoop (push)
    - "Named pipe" QVX connector for Hadoop (pull)
- QlikTech has established partnerships with third-party providers to connect with Big Data sources such as Attivio, DataRoket, and Informatica. A QVX SDK is available to all third-party developers who wish to build custom connectors for any system with an open API.
- QlikView has partnered with Google to provide a visual analytics front end to the Google BigQuery solution.
- In June 2012 QlikTech acquired Expressor Software and now offers the QlikView Expressor Server, which provides a metadata intelligence capability and advanced data integration capabilities.[iv]

Technical Considerations When Working with Hadoop for Analytics

The Hadoop project is an open-source project focused on providing massive computational scalability, however often at the expenses of ease of use and performance. It's worth remembering that Hadoop is not a database, rather it is a framework for distributed computation on a massive data application, and as such does not provide any native mechanism to directly query the data. Each query must be performed by writing a "one-off" program that leverages MapReduce, which is a framework for processing large data sets.

For this reason, other open source projects have added modules to the Hadoop core to try to alleviate the drawbacks that come with extreme scalability. The most famous of these are probably Hive, a data warehouse that sits on top of the Hadoop Distributed File System (HDFS), and MapReduce, which provides an easier way to access Hadoop data through HQL (Hibernate Query Language), a SQL-like query language. Hive provides an easier way to connect to Hadoop data, exposing ODBC (open database connectivity) and JDBC (Java database connectivity) connectors.

Although this is a big step forward in ease of use and integration, Hive is still an incomplete solution for data analysis, especially for non-technical business users. The main issue any non-in memory solution has with Hive is related to the fact that they were not designed for OLTP workloads and do not offer real-time queries or row-level updates. They do not provide an associative experience for business users.[v] In addition, they don't enable organizations to easily meld Big Data with existing enterprise or cloud data. Hive is best used for batch jobs over large sets of append-only data.

Data integration can be another challenge with Hive. Hive is designed to be Hadoop's data warehouse, so it does not accept data coming from other data sources. At the time of writing, a project called SQOOP had recently exited from the incubator stage and is attempting to provide a means to integrate data coming from RDMSs (relational database management systems) with Hive. SQOOP, however, is still quite limited and its stability is unproven. QlikView's ability to natively connect to multiple data sources and define logical connections among them is certainly a more preferred solution to data integration with Hadoop-based Big Data solutions and provides users with QlikView's unique associative experience.

QlikView and Hadoop: A Case Study

King.com is on online gaming property based in Scandinavia. King.com is using QlikView with a Hadoop-based Big Data system to provide business users in the marketing function with rapid insight into customer behaviors captured through their use of the games. Everything from customer browsing activity while on the site to their interactions within each game played and many, many more metrics are captured in an on-premise Hadoop-based deployment. QlikView sits on top of the Hadoop system to provide Business Discovery capabilities to enable King.com to more effectively target new customers, new games, new offers, and so on.

The data volumes acquired are impressive: 1.6 billion new rows are produced each day and stored in the Hadoop system. King.com uses QlikView to provide analysis of aggregated data from the Hadoop system. In their words, having billions of permutations of data reduces the statistical impact of the aggregation effect so that when they are analyzing 211 million rows in QlikView, they have a high degree of confidence that the data is highly representative of the entire data set within Hadoop.

King.com utilizes a single 8 machine cluster to host their Hadoop environment. Each user event is logged, processed, and ultimately made available to QlikView users for analysis via an ODBC connector to Hive. The flow of data from the source gaming systems all the way to the QlikView analytics system is shown in Figure 3.

Figure 3: Flow of data in the King.com implementation of QlikView and Big Data

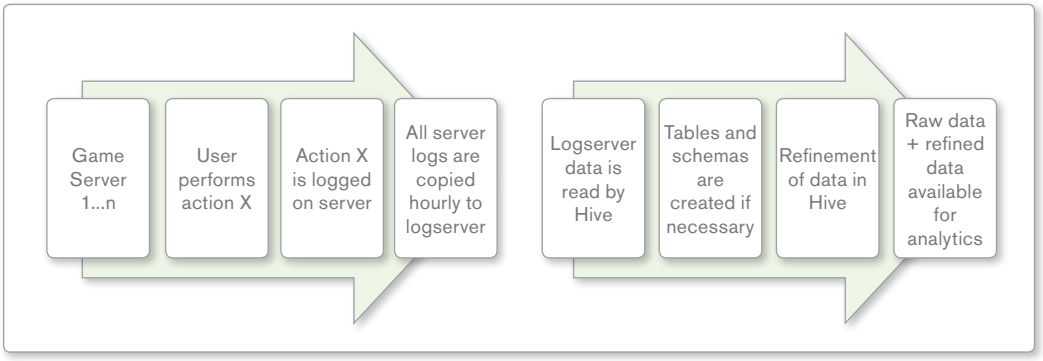
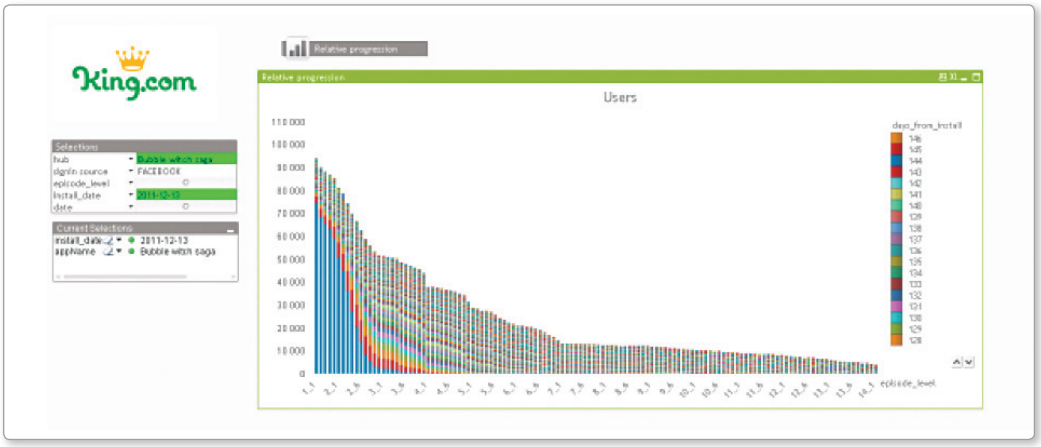


Figure 4: Screenshot of the King.com QlikView application



QlikView and Google BigQuery

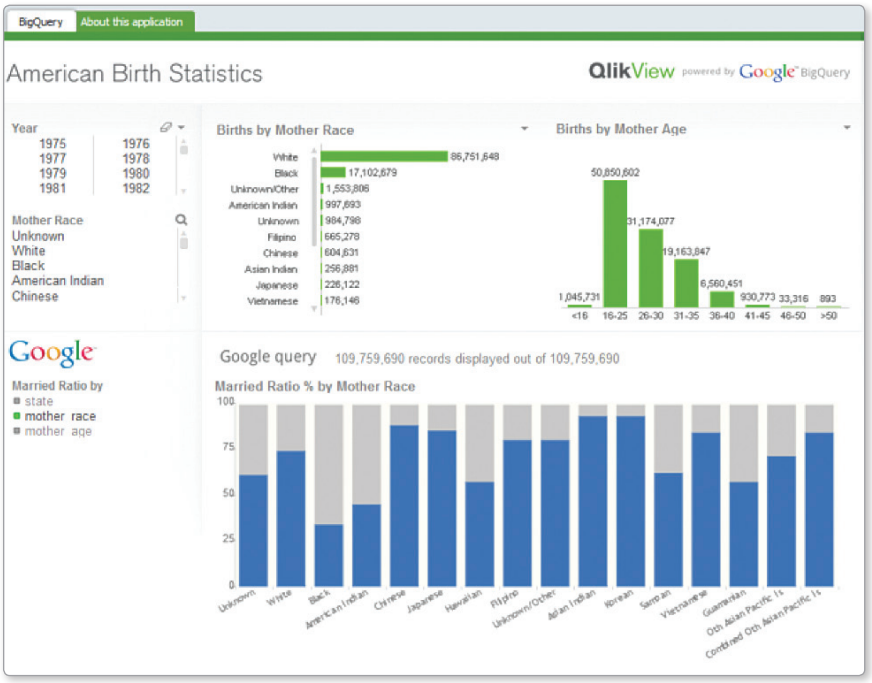
Google BigQuery is a web service that lets business users and developers do interactive analysis of massive datasets – up to billions of rows – without any up-front hardware or software investments. It is scalable and easy to use: BigQuery lets developers and businesses tap into powerful data analytics on demand.

The QlikView Business Discovery platform provides seamless integration to Google BigQuery with its extension and custom connector capabilities. With QlikView's custom connector, users can load BigQuery data into memory and explore information freely rather than being confined to a predefined path of questions. They can remix and reassemble BigQuery data in new views and create new visualizations on the fly for deeper understanding. With QlikView's unique associative experience, business users can navigate and interact with the BigQuery data in any way they want to.

In addition to the custom connector, the QlikView extension object provides a direct connection from QlikView dashboards to Google BigQuery, which enables users to ask ad-hoc questions on large volumes of data and get answers in seconds. It enables business users to ask ad-hoc questions on the BigQuery data that do not exist in-memory and gets instant answers in seconds without writing a single line of SQL.

The QlikView Google BigQuery integration solution enables non-technical and non SQL-savvy users to interact effectively with billions of rows of data in seconds to find what is relevant to them and ask their own questions on the BigQuery data.

Figure 5: QlikView Google BigQuery demo application (http://Qlikview.com/bigquery)



QlikView goes the ‘last mile’ with Big Data

One of the big challenges in telecom is the “last mile” — bringing the telephone, cable, or Internet service to its end point in the home. It is expensive for the service provider to fan out the network from the trunk or backbone – to roll out trucks, dig trenches, and install lines. As a result, in some cases they pass high installation costs down to the end customer — or neglect to go the last mile at all. There is a “last mile” problem in Big Data, too.

Today, most vendors working on the problems of Big Data are focused on processing the data — they are focused on the backbone, to use the telecom analogy. The last mile: this is where QlikTech fits into the picture. QlikTech’s mission is simplifying decisions for everyone, everywhere. With the QlikView Business Discovery platform we have user experience in our DNA. Our business model supports a fan-out to the business users — the corollary of the home in the telecom analogy. QlikView is a great complement to the capabilities of vendors focused on processing Big Data and truly provides that high value, highly relevant component of Big Data, namely providing analytics and meaning to their data, for everyone.

End Notes

[i] According to the Bureau of Transportation Statistics Research and Innovative Technology Administration, in 2012 (ending on the last day of February) there were 9,098,000 departures, compared to 9,125,000 in the same period 2011, a change of -0.3%. For more information see http://www.transtats.bts.gov/.

[ii] On June 13, 2012, eBay's social commerce analyst Palm Norchoovech shared these insights in a presentation titled, "Global Social Analytics @eBay" at the Text Analytics Summit in Boston, Massachusetts. You can find more info here: http://bit.ly/GSnH03.

[iii] For example, QlikView customer King.com sources 1.6B additional records a day into their Big Data solution but extracts only 211M aggregated records for analysis. King.com performs clustering or sampling to choose a statistically relevant subset of the data, and performs analysis just on that.

[iv] To learn more about QlikView Expressor, see this web page: http://www.qlikview.com/us/explore/products/expressor.

[v] QlikView works the way the mind works. QlikView's associative experience enables users to answer un-asked questions. The user's selections are highlighted in green, associated data is highlighted in white, and unassociated data is highlighted in gray. Users can make an unlimited number of selections and all the data in the app — which may have come from multiple source systems — instantly filters around those selections. For more information see the QlikView White Paper, "The Associative Experience" (http://bit.ly/hgf12U).