White Paper

QlikView and Big Data: have it your way

January, 2014

qlik.com

Table of Contents

Executive summary

- Big Data's promised benefits are not realized until there is a way for business users to easily analyze data.
- The key to unlocking value lies in presenting only what is relevant and contextual to the problem at hand.
- QlikView offers two compelling options to handle Big Data: 100% in-memory and Direct Discovery, a hybrid model.
- Either way, customers experience a significant advantage in time-to-insight when it comes to analyzing Big Data.

Introduction

There is an incredible amount of interest in the topic of Big Data at present: for some organizations its use is an operational reality, providing unprecedented ability to store and analyze large volumes of disparate data that are critical to the organization's competitive success. It has enabled people to identify new opportunities and solve problems they haven't been able to solve before.

For other organizations, 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 user-friendly analytics platform accompanying a Big Data solution. It is written for 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.

The two sides of Big Data analytics

Published material on the uses of Big Data usually focuses on running very complex algorithms on massively parallel computing clusters to solve major challenges in academia, government, and the private sector. In fact, the people who create these algorithms are called "data scientists". They have deep experience in math, statistics, data mining, and computer science. Competition for this rare combination of talent is fierce. Given the scarcity of such talent and the time required to program these algorithms, it is natural to consider whether there is an alternative – whether it is possible to harness the power of Big Data analytics for business users.

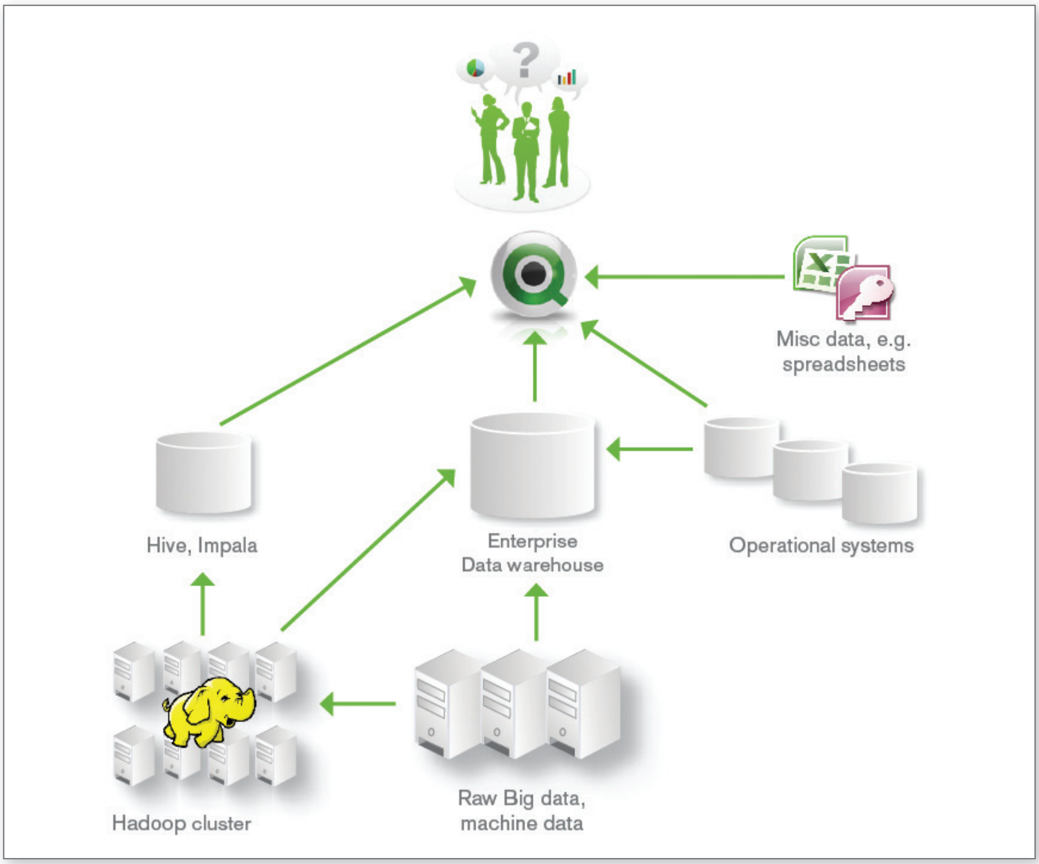
This other side of Big Data analytics is typically performed by business users in a self-service model. Unlike the algorithmic model which seeks to find the needle in the haystack by mining through all the data available, business users are more likely to ask ad hoc questions that result in insights that lead to actionable business decisions such as:

- How have sales of product X performed since we ran the last promotion?
- How effectively is our sales team cross-selling our products?
- How has the supply chain for the manufacturing of product Y been disrupted by a natural disaster?
- Does the transaction history of customer A indicate a pattern of satisfaction?

These types of questions have been posed by business users long before the advent of Big Data, but the inclusion of data sets that did not exist or were not practical to access increases the possibility that the questions are answered to a higher degree of certainty or granularity. In other words, *business users are able to combine their intuition with better data to arrive at more optimal decisions.*

What makes Big Data particularly difficult to work with is that standard relational databases, even if running on the highest-spec computers, cannot process it fast enough. IT managers have turned to several different solutions to solve the Big Data storage and processing problem.

Figure 1: the flow of data from source to analysis


Misc data, e.g. spreadsheets
Hive, Impala
Enterprise Data warehouse
Operational systems
Hadoop cluster
Raw Big data, machine data

©2014 Qlik

How Big Data flows from source to analysis

Figure 1 shows the flow of data from its raw form to a refined form suitable for analysis and presentation. To make an analogy from metal mining, raw ore must be extracted from the earth, transported to plants which use mechanical and chemical processes to refine the metal, and only then can it be fashioned into jewelry or other products.

Likewise, data follows a journey from its raw form to delivering business insight:

- **The fresh, raw data.** The origin of business-oriented Big Data is typically machine data (e.g., server logs, network logs, and RFID logs), transaction data (e.g., website activity, point of sale data from physical stores), and cloud data (e.g., stock ticker prices, social media feeds). This data is often unstructured (strings of text or images) or semi-structured (log data with a timestamp, IP address, and other details). In the common definition of Big Data, this sort of data has high volume (terabytes to petabytes), high velocity (many terabytes of new data per day), and high variety (hundreds of different types of servers and applications each creating information in their own format).
- **The first round of processing.** If cost of storage is the primary concern, the data is often copied to a Hadoop cluster. The Hadoop Distributed File System (HDFS) is an example of a distributed, scalable, and portable file system designed to run on commodity hardware. Hadoop jobs called MapReduce enable highly parallel data manipulation and aggregation, which is often used as a first-level interpretation of the raw data. Apache Hive and Cloudera Impala are part of the Hadoop ecosystem and provide open source means for external systems, such as QlikView, to query the data stored in Hadoop.
- **More processing.** Quite often, organizations will also employ an enterprise data warehouse (EDW) which serves as the central repository for structured data that require analysis. EDWs are designed for not just storage volume but also have robust ETL (extract, transform, load) capabilities hence they play a complementary role with Hadoop clusters. EDWs can extract data directly from the data source, a SAN (storage area network) or NAS (network attached storage) system, or Hadoop clusters. Because data in EDWs is structured and not raw, it is easier to query and represent a higher level of meaning than raw data.
- **The final stage: analysis.** The analysis tool that most closely fits the needs of a typical business user must flexibly integrate data from multiple sources and not make assumptions about where the data comes from or how it is organized. Data modeling must be fast and easily span different data sources. These requirements not only reduce the burden on IT to keep up with business demands, they also empower business users to incorporate additional data in their analysis as need in a timely manner.

Focus on relevance and context

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.

We are seeing a few common misconceptions about Big Data analysis, such as assuming that:

- **The most important data is in the Big Data repository.** Often, the data from the Big Data repository acts as supporting evidence for a discovery initially made in operational data or even in a spreadsheet. For example, a spreadsheet or small database containing customer satisfaction survey results may be the basis for an analytic inquiry, and the data from a Big Data repository allows the user to correlate a customer's customer service or support history with their satisfaction scores.
- **All the data need for analysis is in a single repository.** The process of configuring an enterprise data warehouse to not only copy data from an operational data source but to also perform metadata modeling and transformations could be time consuming or cost prohibitive. It is sometimes better to pull some data straight from the operational source until it is clear that loading it into the data warehouse in a formalized process has enough benefit to the enterprise to warrant the cost and effort.

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., Hadoop repositories, data warehouses, departmental databases, and spreadsheets) in one single, interactive analytics layer.

Relevance: the right information to the right person at the right time

QlikView's approach has always been to understand what business users require from their analysis, rather than to force feed a solution that might not be appropriate. Access to appropriate data at the right time is more valuable to users than access to *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 volumes to one of relevance and value.

Context: what does the big data mean in context of other sources of insight?

The design of QlikView makes it natural to surround data with context. QlikView's associative experience means that every piece of data is dynamically associated with every other piece of data. This means that when a user sees a chart, for example sales by region, that chart is surrounded by interactive list boxes that contain contextual information such as date, location, customer, product, sales history, etc. Any time the user makes a selection in any of the list boxes, all the other list boxes and all other charts are instantly updated with the user's selections. In QlikView, selected values are green, associated values are white, and unassociated values are grey.

For example, if the user selects a particular product in the product list box, a list box containing a list of geographies would show associated regions (where the product was sold) in white and unassociated regions (where the product *wasn't* sold) in grey. This unique capability of QlikView makes it incredibly easy for a business user to focus on a particular product in a particular geography sold to a particular customer and see only the data that is relevant.

The usefulness of the green-white-grey associations is even more apparent working with real business data where there might be hundreds or thousands of products, customers, geographies, etc. Extremely large datasets can be sliced with a few clicks rather than scrolling through thousands of items. Furthermore, representing unassociated values in grey is very powerful. When the user notices that something isn't associated with the current selection when it ought to be, it often points to a potential business problem that needs to be addressed.

With QlikView, context and relevance go hand in hand and quickly take what seems to be a Big Data problem down to something that is quite manageable without any programming or advanced visualization skills.

QlikView offers two approaches to Big Data analytics

Because Big Data is a relative term and the use cases and infrastructure in every organization are different, QlikView offers two approaches to handling Big Data that put the power in the hands of our customer to best manage the inherent tradeoffs between user performance and data volume, variety, and velocity.

100% in-memory for maximum user performance

Despite tremendous advances in hard disk technology and the advent of solid state drives (SSDs), the throughput and latency of RAM vs. disk is still orders of magnitude apart. Therefore, the ideal configuration for split-second responses from QlikView is still to bring all the relevant data needed for analysis in memory.

King.com uses QlikView for Big Data analytics

> “Implementing QlikView has cost less than 20% of the alternative solutions. The payback period was just a few months.”

–Mats-Olov Eriksson,
The main architect of
the analytics system

Background:

- Worldwide leader in casual social games
- Offers 150 games in 14 languages
- 40 million monthly players
- 2 billion rows of new log data per day

Use case:

- Analyze ROI of marketing campaigns
- Track update of new game offers

Technology:

- Logs stored in 14-node Hadoop cluster
- Batch processing create KPIs and aggregates in HIVE
- QlikView connects via ODBC (open database connectivity) to HIVE

For more details download the whitepaper: QlikView Deployment for Big Data Analytics at King.com

QlikView apps can address the amounts of data that are needed to ensure the relevancy of the app for business users. Here’s how:

- **Hardware advances.** Recent trends in large memory available on standard Intel hardware enable QlikView to handle ever-larger volumes of data in memory (which provides users with a super-fast, interactive experience). Also, QlikView distributes the number-crunching calculations across all available processor cores to maximize the performance experienced by the user. Unlike technologies that simply “support” multi-processor hardware, QlikView is optimized to take full advantage of all the power of multi-processor hardware, thereby maximizing performance and the hardware investment.
- **QlikView compression.** Data brought into QlikView’s in-memory engine can typically be compressed to a tenth of its original size, meaning a single 256GB server can handle uncompressed data sets near 2TB in size. QlikView’s compression scheme means the more redundancy in the data values, the greater the compression.
- **Distributed servers.** With distributed servers in a clustered environment, apps can be hosted on different servers. For example, an app containing a smaller amount of aggregated data could be run on a server with less memory while an app with large amounts of detailed data could be configured to run on a larger server, all without the user having to know where the apps are physically hosted.
- **Multi-tiered architecture.** QlikView can be deployed such that one server runs in the background extracting and transforming large amounts of data while another server runs the user-facing app and does not have the added burden of handling back-end tasks. An additional benefit to IT from QlikView’s multi-tiered architecture is that the transactional data source only has to be accessed once. That data can then be reused in multiple QlikView apps without a fresh extract.
- **Incremental load.** Administrators can configure QlikView to load only data that is new or has changed since the last load, thus greatly reducing the bandwidth required from any data source.
- **Document chaining.** Separate QlikView apps, potentially running on different servers, can share selection states, thus a user may be running a small dashboard or summary app and seamlessly switch over to a large app containing detailed data.

Many QlikView customers follow this approach as it satisfies their requirements for access to Big Data while preserving high performance.

QlikView Direct Discovery for truly massive datasets

QlikView Direct Discovery is a hybrid approach that combines the QlikView in-memory data model with external data that is queried on the fly. Tables that fit in-memory are still loaded into memory, but extremely large fact tables are not. The aggregated query result from the external data source is passed back to QlikView, associated with the in-memory data, and presented to the user. The Direct Discovery data set is still part of the associative experience; selections on both the in-memory data and the in-place data are reflected throughout the QlikView app.

By directly querying Big Data sources without a complicated ETL process and associating it with smaller tables from spreadsheets and traditional databases, IT departments are able to open up vast information sources, enhance the data with meaning and context, and present the view to business users who can leverage insights to create more informed strategies and make better decisions.

Furthermore, because data is queried on the fly, users needing to access near-real-time data (to address Big Data velocity issues) can do so. To improve efficiency and reduce the load on the Big Data repository, queried data can be cached. The QlikView administrator is able to customize the length of time cached data is valid to reach a balance between timeliness and user performance.

The hybrid approach of QlikView Direct Discovery alleviates data silos, giving users the data they need when they need it without time or productivity drains. With QlikView Direct Discovery, users gain Big Data access with all of the associative experience of QlikView Business Discovery. Users can continue to explore information freely and generally do not notice the difference between in-memory data and the in-place data. It is important to keep in mind that the performance of Direct Discovery is very much related to the performance of the underlying Big Data repository.

QlikView customers often ask which is better for them: 100% in-memory or Direct Discovery. Our recommendations are below (see Figure 2).

Figure 2: When to Choose 100% In-Memory vs. Direct Discovery

When to Choose 100% In-Memory	When to Choose Direct Discovery
All the necessary (e.g., relevant and contextual) data can fit in server memory. (~200GB compressed)	Data cannot fit in-memory and document chaining is not a viable solution due to analytical requirements.
Users require only aggregated or summary data, i.e. hourly or daily averages, or record-level detail over a limited time period.	Users require access to record-level of detail stored in a large fact table that will not fit in-memory.
Query performance of external source is not satisfactory or the number of queries expected from concurrent users would negatively impact the Big Data repository.	Network bandwidth limitations means that it would take too long to copy raw data to a QlikView server. Direct Discovery queries return aggregated data hence requires less bandwidth.

QlikView and Big Data connectivity

QlikView is designed as an open platform and comes with a number of built-in and third-party connectivity options for Big Data repositories, such as:

- **ODBC for access to data sources with an ODBC interface.** QlikView's ODBC connector connects to most databases due to the wide adoption of the standard by database vendors. For example, Teradata provides an ODBC driver to the Teradata enterprise data warehouse and Cloudera provides an ODBC driver to its popular distribution of Apache Hadoop and Hive, and Impala, its real-time query engine for Hadoop.
- **QVX for access to non-standard sources.** QlikView provides an open data protocol (QVX, or QlikView Exchange) which can be leveraged by developers to connect to custom data sources via two different methods:
    - **Disk-based QVX file extracts (push)**
    - **"Named pipe" QVX connector (pull)**
- A QVX SDK (software development kit) is available to all third-party developers who wish to build custom connectors for any system with an open API.
- **QlikView Expressor.** In June 2012 Qlik acquired Expressor Software and we now offer the QlikView Expressor Server, which provides a metadata intelligence capability and advanced data integration capabilities. [1]
- **Partnerships.** Qlik has established partnerships with third-party providers such as Attivio, DataRoket, and Informatica to connect with other Big Data sources. QlikView has also partnered with Google to build a custom connector, providing a visual analytics front end to the cloud-based Google BigQuery solution.

Qlik 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 telecom providers pass high installation costs down to the customer — or neglect to go the last mile at all.

There is a "last mile" problem in Big Data, too. Today, most technology providers working on the problems of Big Data are focused on processing the data — they are focused on the backbone, to use the telecom analogy (or the plant, in the ore mining analogy).

The last mile is where Qlik fits into the picture. Qlik'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 provides a high-value component of the Big Data solution: analytics for everyone.

[1] To learn more about QlikView Expressor, see this web page: qlik.com/us/explore/products/expressor.

qlik.com