



“BI”g Data –

How Business Intelligence and  
Big Data Work Together

By BI & AI SIG in ODPi

[odpi.org](http://odpi.org)



# Overview of BI & AI SIG

Technology around Business Intelligence (BI) has been developed since the 1990s. Most of the BI tools stem from Relational Database Management System (RDBMS). After many years of development and evolution, BI is now a pretty mature area that most modern corporations embrace. Big Data technology use continue to grow, and organizations have grown their investments in this technology in the hope of gaining more business value. Although there are a lot of fundamental differences between traditional RDBMS and Big Data, they also have a lot of similarities. Moreover, it's almost impossible for existing BI vendors to ignore RDBMS to dive completely into Hadoop/Big Data. Hence, considering both in the roadmap is inevitable.

Exploring an optimal BI approach to help users consuming data effectively is one of the critical step in analytics, which aligns perfectly with ODPi mission of simplification and standardization of the big data ecosystem. Hence, this project is chartered to bridge the gap so that BI Tools can sit harmoniously on top of Big Data and RDBMS, yet provide the same, or even more, business insight to the BI user who also has Big Data in the backend. From a BI vendor perspective, this project aims to find an effective way for connecting and querying Big Data without reinventing the wheel. From a BI user perspective, this project should help to provide an objective guideline for evaluating the effectiveness of a BI solution, and/or other related “middleman” technologies such as Apache Hadoop, Apache Hive, Apache Drill, Presto, Apache Impala, Apache Phoenix, etc. After socializing this idea with 5 of the most prominent BI vendors in the industry, we decided to join force within this project hosted by ODPi to publish a white paper and share with the industry what can we learn from each other in the aspect of putting BI on top of Big Data.

Finally, if you have followed this group, you should know that this project was previously called BI & Data Science SIG. However, because of the fast-growing AI usage, AI models can now run on Big Data much more cost effectively than years ago. This results in fostering the convergence of traditional BI and AI disciplines (Machine Learning, Deep Learning... etc). Hence, we decided to align our focus to the industry trend by changing our name to ODPi BI & AI and become a full fledged ODPi project to better explore how this phenomenon is going to impact the Big Data ecosystem.



# Introduction

It doesn't matter how much data you have; unless you can get the insight from it, it is just bits and bytes occupying the storage.

On the other hand, even if you have an excellent tool to present insightful visualization, unless you have good amount of quality data to support, it's just dog and pony show. Therefore, it's important to know how to combine BI with Data, especially Big Data, to gain true insightful business value. That is also the goal of this article – to explore what is the best way for putting BI on top of Big Data effectively. We are honored to have five BI Leaders to participate in this publication: IBM, MicroStrategy, Qlik, SAS, Tableau, in alphabetical order. Even though their opinions do not represent ODPi, their feedback will give us valuable insight into the industry trend of how BI comes along with Big Data.

BI & AI SIG welcomes everyone, individual or company, who has the same passion to join the group and contribute to the Big Data industry. Please feel free to contact us at [odpi-sig-bi@lists.odpi.org](mailto:odpi-sig-bi@lists.odpi.org) if you have any question.

## What is the preferred BI/SQL Connector (Hive, Presto, Impala...etc) for your BI Tool to connect to Hadoop? Why?

 Recent versions of IBM Cognos Analytics support the following SQL interfaces to Hadoop:

- IBM Big SQL
- Apache Impala
- Apache Hive
- HPE Vertica
- Apache HAWQ
- Presto
- Apache Spark SQL



IBM Cognos software generates SQL tailored for the type and version of each of these technologies. Cognos optimizes queries to minimize user wait time by delegating data computations to the Hadoop cluster as much as possible.

For certain nascent technologies, such as early versions of Apache Hive, the SQL support may be limited and common BI paradigms, such as widowed aggregates, may not be supported. This necessitates the Cognos software to perform those computations itself on larger volumes of data than the user sees on their screen.

In robust, mature technologies like IBM Big SQL, there is high fidelity between what it supports and what is required by the types of analyses one can do in Cognos. With these technologies, virtually all the data processing is performed in parallel by the nodes of the Hadoop cluster. Usually, only as many rows as there are points to be plotted in a Cognos dashboard visualization need to be transferred from Hadoop to Cognos. This practice is preferred.

**MicroStrategy** Hadoop has become the scalable platform for storing and processing large amounts of data.

It has found widespread applications in enterprise where SQL already represents the de facto language for data analysis. This combination has led to the development of a variety of SQL-on-Hadoop systems to target a similar set of analytical workloads. In the Hadoop ecosystem, you can store your data in one of the storage managers (for example, HDFS, Apache HBase, Apache Solr, etc.) and then use a processing framework to process the stored data. While the various SQL-on-Hadoop systems target the same class of analytical workloads, their different architectures and design decisions impact query performance and concurrency.

MicroStrategy has a vast library of native gateways and drivers, that allows a user to connect to any enterprise resource, so you can fully leverage existing investments—from databases, enterprise directories to cloud applications, physical access control systems and various SQL on Hadoop Technology. Some of the more popular ones that have been adopted by our customers are SparkSQL, Impala, Apache Drill, Apache Phoenix, Presto, Apache Druid in addition to a long tail of SQL Engines such as IBM BigSQL, Oracle BigData SQL, Polybase, et al. In addition to supporting more complex OLAP workloads on Hadoop, MicroStrategy partners with “middleware” companies such as Arcadia Data, AtScale, Jethro, and Kylogence, among others.

Besides leveraging all SQL-on-Hadoop technologies, MicroStrategy has invested in a Native Hadoop Gateway to address the workflow where business users want the flexibility to browse the HDFS and work with native file formats, such as parquet, ORC, Avro, and JSON. The need to build a native driver addressed two key workload limitations of current SQL on Hadoop’s ODBC/JDBC drivers – The first was the transfer rate of data on SQL on Hadoop drivers are slow and at very large volumes, as it can take hours to build an OLAP or in-memory cube for large scale analytics. The second limitation was that SQL-on-Hadoop drivers don’t support Data Preparation capabilities for business users to wrangle the data. The MicroStrategy’s native Hadoop driver uses Spark as the execution engine and does parallel transfer of data to speed up analytics workflows.



Selecting the right SQL-on-Hadoop technology requires a detailed analysis. Richness of SQL, performance of joins, storage formats, user defined functions, multi-user workloads, data federation, and the capability to run on non-Hadoop clusters are just a few factors to consider. Choosing the right data infrastructure and architecture streamlines your workflows, reduces costs, and scales your data analysis. What makes the choice more difficult is that almost all SQL-on-Hadoop vendors make bold claims on speed and how their engine fares against the TPCB benchmarks.

MicroStrategy's platform supports various workloads across all the SQL-on-Hadoop engines and have optimized the user experience with a complimentary in-memory technology. The MicroStrategy Platform provides the optimal choice and breadth of SQL richness according to the workload – batch, interactive, in-memory. Users can pick and choose the best of breed based on the type of analytics, workloads, and SLA requirements.

**Qlik**  Qlik has validated connectivity with the following Big Data - Apache systems with built-in connectors:

- Hive, Hive w/Apache Tez, Hive w/ LLAP
- Impala
- Presto
- Drill
- SparkSQL
- Phoenix (HBase)
- Qlik has also validated connectivity with the following Big Data SQL systems:
- Apache Kylin/Kyligence
- Jethrodata
- AtScale
- Kyvos

These systems all use different types of processing engines and techniques to extract data from Hadoop, and Qlik can leverage any of these technologies effectively. We do not have a preferred solution as Qlik works with all the different Hadoop distributions, and each has their own preferred connectivity method.

Some customers will also leverage additional tools (Jethro, AtScale, etc) to boost performance even more – and so we work with those as partner technologies as well.

**SAS**  At SAS, we believe that the connectivity approach depends on the situation.

Our SAS/ACCESS® software provides capabilities which far exceed those of the typical “Connector.” SAS/ACCESS® products are designed to:



- Enable SAS users to code using the SAS language and transparently access database objects
- Push WHERE clauses to the database
- Transparently push JOINS to the database
- Push SAS procedure processing to the database
- Load data into the database using bulk loading tools provided by the database vendor
- Enable SAS function calls to be converted to database functions
- Enable SAS users to customize CREATE TABLE statements generated by SAS
- Handle localization and international issues

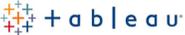
The starting point for accessing Hadoop is SAS/ACCESS® Interface to Hadoop. SAS/ACCESS Interface to Hadoop connects to Hive and is available for both the SAS Viya and SAS® 9 platforms. This interface supports many commercial versions of Hadoop. SAS recommends this approach because Hadoop/Hive is the preferred access method for our customers. Hive is pervasive.

SAS/ACCESS® Interface to Hadoop provides all the capabilities listed above. Bulk loading is implemented as a direct write to HDFS via WebHDFS or the HDFS API. Certain classes of SELECT statements can use this same mechanism because it greatly enhances read performance.

If your preferred BI architecture includes Impala, then SAS/ACCESS® Interface to Impala can be added to the mix.

SAS/ACCESS® Interface to ODBC can be used for situations where SAS does not have a dedicated SAS/ACCESS® product for a specific SQL engine (for example, Presto). This is a general use SAS/ACCESS interface that can connect to any ODBC 3.5 (or later) compatible driver.

Note that SAS® In-Database products take the idea of a connector to a whole new level. SAS® In-Database products run SAS code inside the Hadoop and databases greatly increasing performance and minimizing data movement.

 **Tableau is on a mission to help people see and understand their data.**

To achieve this mission, Tableau promotes the democratization of data with the fundamental belief that the people who know the data the best should be empowered to ask questions of their data.

The approach to visualizing Big Data stored in Hadoop is focused on three themes:

- Connectivity - providing access to all data regardless of size, how it is stored, whether it is structured or unstructured in format
- Performance - delivering fast interaction with all data
- Discovery - finding the right data

Tableau with Hadoop makes this possible. Tableau interfaces with Hadoop using SQL. The SQL that Tableau generates is standardized to the ANSI SQL-92 standard.

## Tableau connects with Hive & Spark SQL but optimally with Impala

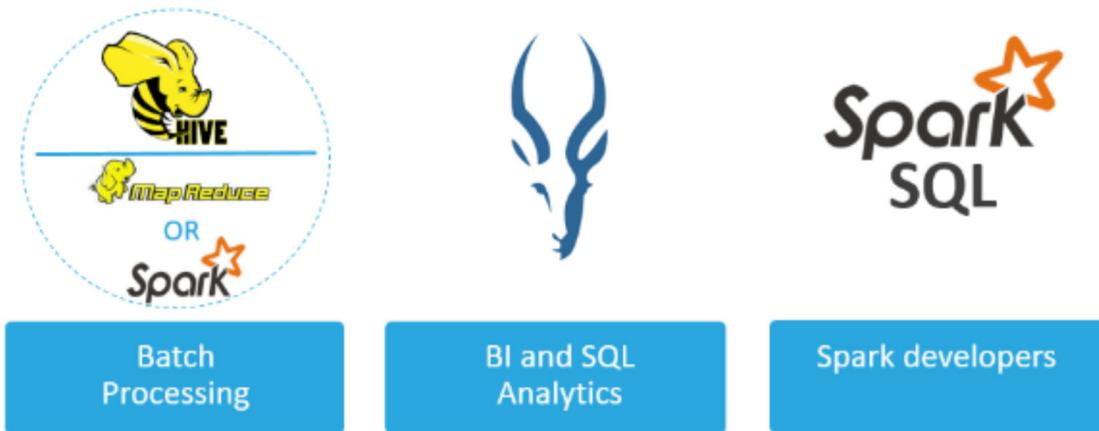


Tableau works with Hive and Spark SQL. It works optimally with Impala. Tableau exposes a number of unique capabilities for connections with Hadoop Hive. These capabilities include:

- XML Processing - Tableau provides a number of user-defined functions (UDFs) for processing XML data using XPath. These functions provide users with the ability to extract content, perform simple analysis, and filter the XML data.
- Web and Text Processing - In addition to the XPath operators, the Hive query language offers several ways to work with common web elements and text data including:
  - JSON Objects - Retrieve data elements from strings containing JSON objects.
  - URLs - Extract components of a URL such as the protocol type or host name, or retrieve the value associated with a given query key in a key/value parameter list.
  - Text Data - Find and replace text in Hive from within Tableau.
- On-the-Fly ETL - Custom SQL enables users to define their data connections using complex join conditions, pre-filtering, and pre-aggregation.
- Initial SQL - Initial SQL allows a user to specify a collection of SQL statements to be performed immediately after a data connection is established. Typically this is done to tune performance characteristics or develop custom data processing logic.
- Custom Analysis with UDFs and MapReduce - Tableau enables users to implement UDFs, user-defined aggregate functions (UDAFs) and arbitrary SQL expressions from Hive by using “pass through” functions. These functions are typically built as Java archive (JAR) files that can be copied across the Hadoop cluster. Users can also exercise explicit control over the execution of MapReduce operations in Custom SQL.



Hive acts as a SQL-Hadoop translation layer, translating the query into MapReduce, which is then run over HDFS data. Hive is the best for batch.

Impala executes the SQL statement directly on HDFS data (bypassing the need for MapReduce). Impala is purpose built for interactive BI on Hadoop.

Tableau also supports Spark SQL, an open source processing engine for big data that can perform up to 100x faster than MapReduce by running in-memory rather than on-disk. Spark SQL in the future will enable Spark developers to inline SQL as steps within their Spark application.

# Best practices for BI tool to connect to both Hadoop and RDBMS

**IBM** As Business Intelligence continues to shift away from static reports executed in batch overnight towards highly interactive ad hoc analysis, the emphasis of best practices is to minimize the time a user waits to see the data.

The Cognos Analytics user experience and best practices for reporting on Hadoop data is identical to reporting on traditional data warehouse technologies. In both cases the obvious but most important principle is that all else equal, less is faster. In practical terms, this means:

- Filter as much as possible to limit the processed data to only what's needed
- Avoid unnecessary complexity, particularly in calculations
- Avoid unnecessary data type conversions

There are other best practices for getting the performance possible in Chapter 6 of the IBM Cognos Dynamic Query Redbooks publication which are equally applicable to data stored in Hadoop as well as data stored in other systems.

When blending data from Hadoop and an RDBMS together, IBM Cognos Analytics employs a filter join optimization that reduces the amount of data that needs to be retrieved from Hadoop. The query to the RDBMS is executed, and a set of key values is gathered and then added to the query that is executed against Hadoop. By extending the predicates (filter criteria) sent to Hadoop, the amount of data processing on the Cognos server needed by the join is reduced. As a result, performance can be improved by several orders of magnitude. More information about this optimization can be found here: [https://www.ibm.com/support/knowledgecenter/en/SSEP7J\\_11.0.0/com.ibm.swg.ba.cognos.ug\\_fm.doc/t\\_dqm\\_join\\_opt.html](https://www.ibm.com/support/knowledgecenter/en/SSEP7J_11.0.0/com.ibm.swg.ba.cognos.ug_fm.doc/t_dqm_join_opt.html)

If at any point live connections to the data store aren't performing adequately, a Cognos Analytics user can easily grab a slice of the data to be stored in Cognos and run their analyses on that.

**MicroStrategy** As the data landscape evolves and the Hadoop-related technologies improve, it is now possible to build interactive BI applications on top of Hadoop sources.

In the early days of Hadoop, the MapReduce based processing was mostly dedicated to long-running, batch processing workloads. With the advent of SQL-on-Hadoop technologies, the adoption of Hadoop-based systems rapidly increased as it opened opportunities for use cases that required the use of existing technologies originally built for RDBMS systems.



Furthermore, open-source SQL-on-Hadoop engines have also significantly improved over the last couple of years in terms of performance, with a lot more reasonable response times that have enabled interactive applications, including BI use cases. On the commercial side, data platforms have also taken advantage of native implementations, proprietary file formats, and indexing/caching techniques to significantly reduce latency and increase concurrency.

MicroStrategy enables customers to take advantage of their data investments by providing a wide range of data connectors to all leading SQL-on-Hadoop platforms in the market as well as commercial adjacent technologies that aim at optimizing data workloads on Hadoop systems and data federation platforms.

Many successful organizations that have been long-term users of BI platforms such as MicroStrategy, have followed an approach where the data lake plays a complementary role with their traditional Enterprise Data Warehouse. The agility and flexibility of the data lake allow these organizations to do data exploration over rapidly changing data sources, as well as rapid application prototyping. However, when it comes to operationalizing an enterprise analytics application, the enterprise data warehouse is still a very valuable asset for organizations as it provides a governed business context to data.

This does not mean that organizations have to stick to traditional relational sources. SQL-on-Hadoop platforms allow them to model data, on either materialized or virtualized tables, so that a governed semantic layer can be modeled on top of it.

The success and adoption of enterprise analytics applications depends on how much users can trust the data and the insights derived from it. Any analytics application, whether it's an operational dashboard deployed to thousands of users or a self-service application available to a select group of data-savvy users will greatly benefit from a unified, governed data model.

A governed model not only ensures consistency in the business definitions of data (i.e., single version of the truth), it provides the framework to enforce security around access to data and privileges to perform actions related to the data. For example, analytics and application architects can configure the application so that users are only allowed to see data within their purview. Likewise, admins can enable or limit users from performing actions on the data such as sharing, exporting, or creating new reports.

A governed semantic layer also enables the reusability of the business definitions of data across the organization. The benefit of object definition reusability has two angles. The first is increased efficiency in application development as multiple applications can be built on top of a solid foundation, eliminating the need to create things repeatedly and in silos. The second is the increased level of trust by providing a semantic layer of certified definitions. For example, a business entity (e.g., "Customer" or "Employee") has a unique definition, including a consistent source that ensures a unified view across applications.

Finally, the MicroStrategy Platform gives the ability to create an enterprise governed model that provides business context to data across multiple data sources, whether traditional relational sources or Hadoop-based sources. Also, the MicroStrategy security model allows organizations to leverage security policies that exist in the data sources by connecting and executing queries under the specific identities of end users, thus enforcing data-level security defined by such policies.

**Qlik**  Our best practice is to leverage what Qlik does best, extract the needed data from all systems and associate it in memory.

Qlik's architecture is unique in that we aren't dependent on any one system or technology for integration; we have our own built in ELT (extract-load-transform) engine that feeds our in-memory columnar engine.

## The Associative Difference™

Qlik's Associative Engine was built specifically for interactive, free-form exploration and analysis



- ✓ **All your data**
  - Brings together many different data sources without complex modeling
  - Indexes all your data to find all the possible associations
  - Leaves no data behind
- ✓ **Explore without boundaries**
  - Explore, search and pivot based on what you see
  - Instantly updates analytics and highlights associations based on interactions
  - No boundaries or restrictions
- ✓ **Speed of thought**
  - Powerful on-the-fly calculation and aggregation for large numbers of users
  - Seamlessly handles both big and small data

**SAS**  SAS believes in using the best tool for the specific task at hand.

In BI, traditional Hive is seldom the best choice, although it can work. SAS recommends using Apache projects that are designed with BI workloads in mind: Hive Live Long and Process (LLAP) and Impala. Both products provide better concurrency and performance than tradition Hive processing.

The recommended approach for connecting to Apache Impala is SAS/ACCESS® Interface to Impala.

The recommended approach to connecting to Hive LLAP is SAS/ACCESS® Interface to Hadoop.

For other connection needs, SAS/ACCESS® Interface to ODBC is a viable alternative.

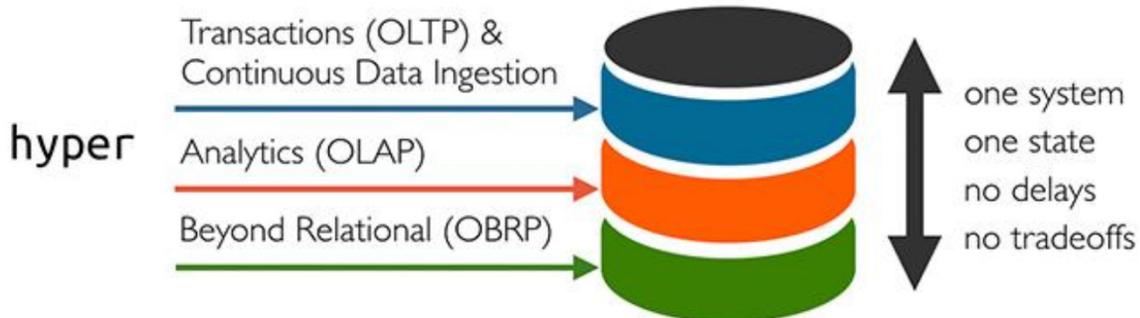
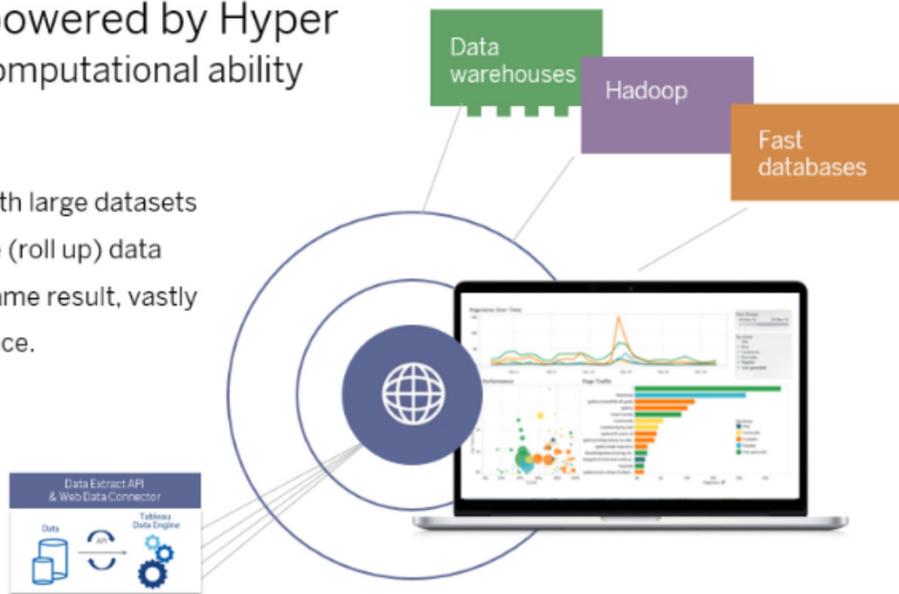
 Tableau is based on a hybrid data architecture that enables live querying of data or extracting data into Tableau’s in-memory columnar data engine called Hyper. Hyper is a fully in-memory columnar data store technology.

Tableau uses Hyper to federate the data from Hadoop and RDBMSs by joining and extracting the data from Hadoop and any RDBMS into Hyper on a scheduled basis. Tableau connects to live data from an RDBMS using a native database driver, connects to data from Hadoop using a variety of methods (Apache Drill, Hive, Impala, Presto, or Spark SQL), and federates it in Hyper.

Benefits of Hyper include up to 5X faster query performance with large or complex datasets, fresher data with up to 3X faster extract creation, and support for large data extracts (up to billions of records).

## Data engine powered by Hyper Extremely fast computational ability

- Fast performance with large datasets
- Optionally aggregate (roll up) data before extracting. Same result, vastly improved performance.



# Recommended BI architecture to query data in Hadoop

**IBM** Any deployment architecture of Cognos Analytics is well suited to query data in Hadoop, since in all configurations the Cognos software will delegate the data processing to the Hadoop cluster as much as possible ensuring only minimized result sets of data need to be retrieved.

**MicroStrategy** In order to provide a wide range of options to address the various uses cases related to “Big Data” analytics, MicroStrategy provides two methods for connecting and querying data from Hadoop-based systems.

The most commonly used approach is to connect via any of the provided connectors for SQL-on-Hadoop data engines such as Hive, Impala or SparkSQL to name a few. With this approach, the MicroStrategy engine connects and queries the Hadoop source via an ODBC or JDBC connector. The second method is via a native connector to HDFS, the MicroStrategy Hadoop Gateway.

The first approach leverages the fact that data stored in HDFS is exposed by the SQL-on-Hadoop interfaces as tables, just like any traditional database. Data is exposed as consistent table definitions which in turn can be modeled and queried by analytics tool such as MicroStrategy. From the analytics tools perspective, connecting and querying data becomes a familiar, well solved problem. Available tables, including their definition in terms of column names and data types, can be easily discovered by standard database catalogue queries. Users can seamlessly create a semantic layer for the data that provides business context that end users can understand and use. MicroStrategy also provides connectors for query accelerator engines such as Jethro Data, which provide an additional layer of indexing and caching that speed up the performance of interactive queries which are common in BI applications such as interactive dashboards or ad hoc reporting.

The second approach employs a native connector provided by MicroStrategy that connects and fetches data directly from the underlying Hadoop Distributed File System or HDFS. This native connector, also known as the MicroStrategy Hadoop Gateway, is deployed directly on an edge node of the Hadoop cluster and leverages the distributed Spark engine running on the nodes of the cluster. The MicroStrategy engine connects and fetches data directly from the cluster without the need to go through Hive or any other intermediate layer. This approach overcomes the throughput limitations of ODBC/JDBC-based connections by fetching and transferring data in parallel from the nodes of the cluster into the MicroStrategy server.



These two methods are complementary. One is not a complete replacement for the other. The native Hadoop Gateway requires users to know what files in the HDFS system contain the data of interest. It doesn't have the advantage of using a table catalogue like SQL-on-Hadoop engines. On the flip side, users can directly explore and analyze the data that exists in Hadoop without requiring the administrator to model data in Hive first. The Hadoop Gateway offers a great way to enable users with agile data discovery and rapid application development and prototyping as it allows them to query extracts of data via filtering and aggregation – all from a self-service user experience.

Users may choose to populate in-memory cubes with extracts of data, which they can analyze efficiently using the standard reporting and visualization tools with instantaneous response time. Alternatively, users may opt to work with a live connection to the data source in order to have access to the full depth of the data, understanding that they will incur a performance cost, because every query is performed live on the source.

Once users have identified the data that they need to create a governed data model, they may opt to operationalize the application using a SQL-on-Hadoop connector after the data has been also modeled in the intermediate layer as tables. They may either opt to create governed, certified datasets that can fetch data on a scheduled basis to maintain in-memory data for maximum performance or they may opt to use live connections to the data source where queries are executed live.

The native Hadoop Gateway connector is a better alternative if users require to perform extensive data wrangling or data transformations on the source data before publishing an in-memory dataset. This is because the Hadoop Gateway is able to perform data transformations on the cluster using native Spark functions which can be processed very efficiently at scale.

**Qlik**  Best practices here depend on how the customer prefers to run.

Qlik can run on-prem, cloud, or in a hybrid configuration to support a customer's topology. We offer preconfigured servers on Azure, AWS, and GCP currently. Recommendations involve basics, such as: for faster extraction from a Hadoop system, place Qlik servers on edge nodes on the same network (better bandwidth is always better). Data volumes are larger, so more memory and compute is often required – a 24 core 512GB RAM box will usually support several thousand users depending on concurrency. I have attached a copy of our scalability guide for reference.

**SAS**  To query data in Hadoop in a BI architecture, SAS' recommends SAS/ACCESS® Interface to Hadoop combined with Hive running on the Hadoop cluster.

The SAS/ACCESS Interface to Hadoop provides many capabilities that make it easier for SAS users to interact with their Hadoop environments. Here are some of the capabilities:

- Enable SAS users to code using the SAS language and transparently access database objects



- Push WHERE clauses to the database
- Transparently push JOINS to the database
- Push SAS procedure processing to the database
- Load data into the database using the HDFS APIs
- Enable SAS function calls to be converted to database functions
- Enable SAS users to customize CREATE TABLE statements generated by SAS
- Handle localization and international issues

SAS is unique in that it can act as a federation layer and combine data from many disparate data sources. SAS/ACCESS® software can combine data from Hadoop with data from many relational databases. A SAS user would connect to both Hadoop and, for example, PostgreSQL. Then the user issues queries, which return subsets of the data. SAS could then do the join.

SAS provides in-memory analytics servers (SAS® Cloud Analytics Server and the SAS® LASR Server). Using SAS/ACCESS® software and SAS® In-Database technologies, our users can pass data between Hadoop and these in-memory servers. As discussed above, SAS orchestrates processing so that data is processed where it lives. This enables SAS to pull preprocessed subsets of data into the in-memory environments resulting in highly performant advanced analytics processing.

It is common for the Hadoop data to be much larger than the RDBMS data. In this situation, the smaller RDBMS result set can be loaded from SAS into Hadoop, and the join performed on the cluster. As we discussed previously, SAS tries to push as much processing into the database possible. This can significantly limit the amount of data that must be moved. This practice is not unique to Hadoop. It can be applied to any data source supported by SAS.

Many customers prefer to access their Hadoop environments using a non-Hive approach such as Impala. For these situations, SAS/ACCESS® Interface to Impala can be added to the mix. Many of our customers use SAS/ACCESS® Interface to ODBC to access their Hadoop clusters. SAS is very flexible.



For applications requiring performance and speed of query, Tableau works with Apache Drill, Impala, Presto, and Spark SQL.

Data can be federated from Hadoop and relational databases into Hyper for data sets up to a billions of records. Tableau provides the flexibility for customers to access data from Hadoop, relational databases, or Hyper directly or federated in Hyper.

# How does BI running advanced analytics including Machine Learn algorithm on Hadoop?

**IBM** The intent is to enable the advanced analytics to complete as quickly as possible while still providing accurate, actionable information to the user.

When the row count of a result set is below 10K, the processing is done on the full result set. If it exceeds that threshold CA requests a statistically valid sample of the data – Bernoulli samples for some technologies that support it like IBM Big SQL, and systematic sampling (basically taking every Nth row) for other technologies.

When Cognos Analytics has sampled data as input to an analysis, there will be a message in the “warning” icon of a visualization indicating that the results are based on sampled data to improve performance.

**MicroStrategy** Today is an exciting time for data scientists: increased availability of Big Data platforms and cheaper computing resources have made feasible a tremendous variety of use cases across every industry.

Despite this, data scientists’ jobs have not gotten easier: greater volume and variety of data, complicated by implementation constraints that vary from project to project, have made it harder to see their results in production. In fact, this ability to bridge the gap – from research project to operational solution – is what separates companies who have evolved into AI driven enterprises from those who have not.

So, how should data scientists think about the design of their solutions and the data that fuels them?

One answer is to begin with the end in mind. Ask, “what is the specific outcome we need from this data science project?” Is the team being asked to build a dashboard to display their predictions, or integrate them inside the workflows of another application? Do models need to be updated throughout the day, or is a nightly refresh acceptable? Understanding and planning for the deployment requirements at the beginning of a project helps data scientists frame the entire project and envision its eventual impact on the organization.

This goes without saying: data has a vital role. Today, we’re awash in data, but it’s spread across several databases and data platforms. Some are in the cloud and others are on-premises. And, how consistently are enterprise



security and data privacy policies applied (and complied with) by data scientists? The reality is that there's rarely a single, secure source system that provides all of the training data.

Data scientists turn to Python, R, Spark, or some combination of these and other tools to provide a consistent solution for extracting data from these sources in addition to wrangling, visualizing, and training machine learning models. This lets them pick the right tool for the job (and it could change from one project to the next). This need for flexibility might come as a surprise to IT and internal governance teams. Can't data scientists choose one? That flexibility could be meaningful to the bottom line: a 1% gain in predictive accuracy or 1% faster model could lead to millions of dollars gained or saved. Data scientists' success, and the realization of the business benefits, rests on the availability of an open analytics environment with few barriers to experimentation.

It feels like we've been here before. In the 1990s, we saw the rise of Business Intelligence (BI) systems. BI made raw data available for analysis by less technically oriented, but acutely knowledgeable business users. They got the freedom to explore and analyze data in a centralized environment, and IT got a layer of governance and a scalable platform. BI laid the groundwork for the transformation to a data-driven organization.

So, what's different about data science? What's preventing data scientists from making the leap into production? Can BI help?

Where teams tend to struggle is in the project's last mile: reaching people who can do something with their insights. It's also expensive to build customized solutions for individual projects that reach these end users where they want to be reached. Luckily, BI systems can facilitate the last mile: the workforce can act on predictive insights that are stored in traditional RDBMS, Big Data, data sets pushed directly into the BI system using APIs, or even predictions that are generated inside the system. With the right BI tool, a broad spectrum of data science solutions is possible on one platform: dashboards, reports, embedded analytics, and entire end-to-end applications.

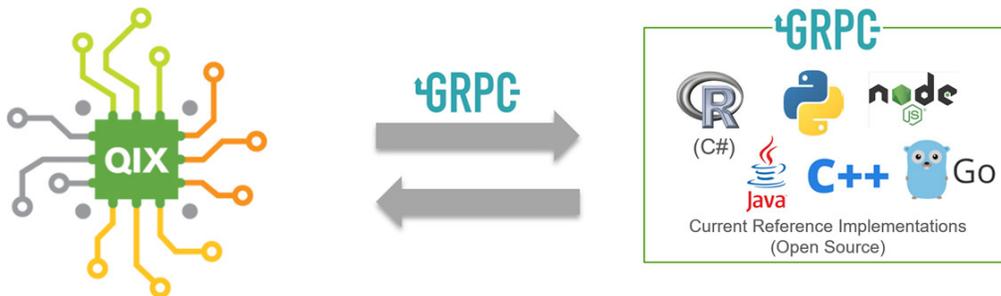
At its core, MicroStrategy is an open data platform, providing secure and governed data that can be queried by data scientists through easy-to-use APIs. These enable teams to develop predictive workflows on top of a solid foundation of trusted data. Likewise, the outputs of predictive workflows can be operationalized on the BI stack in the form certified datasets. From here, insights can be accessed and acted upon by the rest of the organization.

In a future, broader role, BI could become a centralized data inventory. It could provide access to hundreds of trusted sources in a secure, scalable fashion while maintaining the openness and flexibility required by different user groups. For data scientists in particular, imagine what could be possible if they had access to a centralized, trusted enterprise warehouse that didn't require them to create individual solutions for each source system? Could this be the next wave for BI?

## Qlik Qlik has taken a novel approach to incorporating ML/AI into our engine.

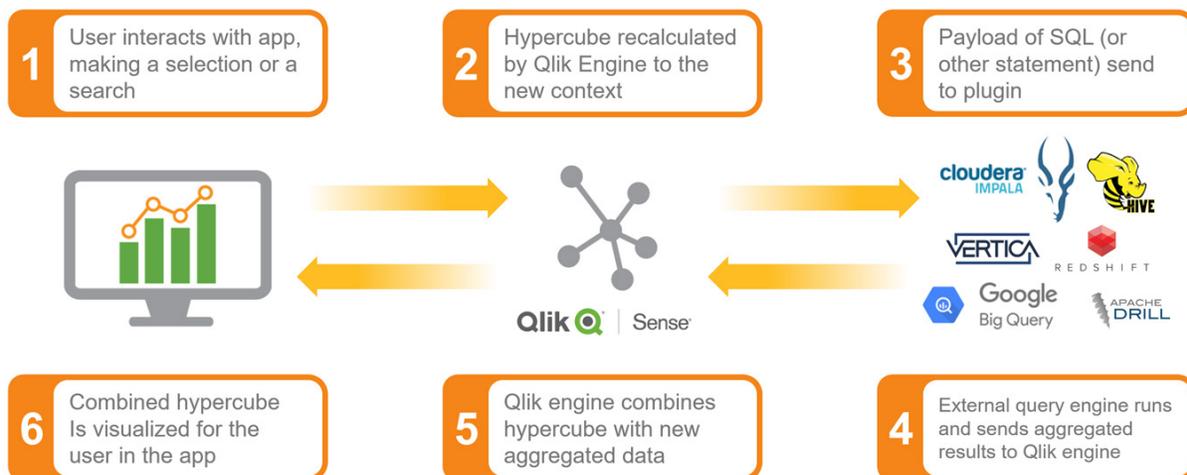
We have added AI for helping users quickly analyze data with our new Cognitive engine in June 2018 release. However, for true ML/AI we use a server side technology (not desktop) called Advanced Analytic Integration (AAI). This technology uses our API to interactively send data to third-party engines for processing calculations. That data is then returned via API to Qlik and visualized.

### What makes our AAI approach unique Qlik's Associative Engine working with all third party engines



- Connectors can be built for any third party engines, through open APIs
- As the user explores, only a small set of chosen and relevant data is sent
- Results are instantly visualized for the user, allowing for further exploration
- Third party engines quickly process smaller, user-specific data sets
- Far more speed than conventional batch techniques
- Results for each user are sent back to Qlik Sense in real-time

<https://github.com/qlik-oss/server-side-extension/tree/master/examples>

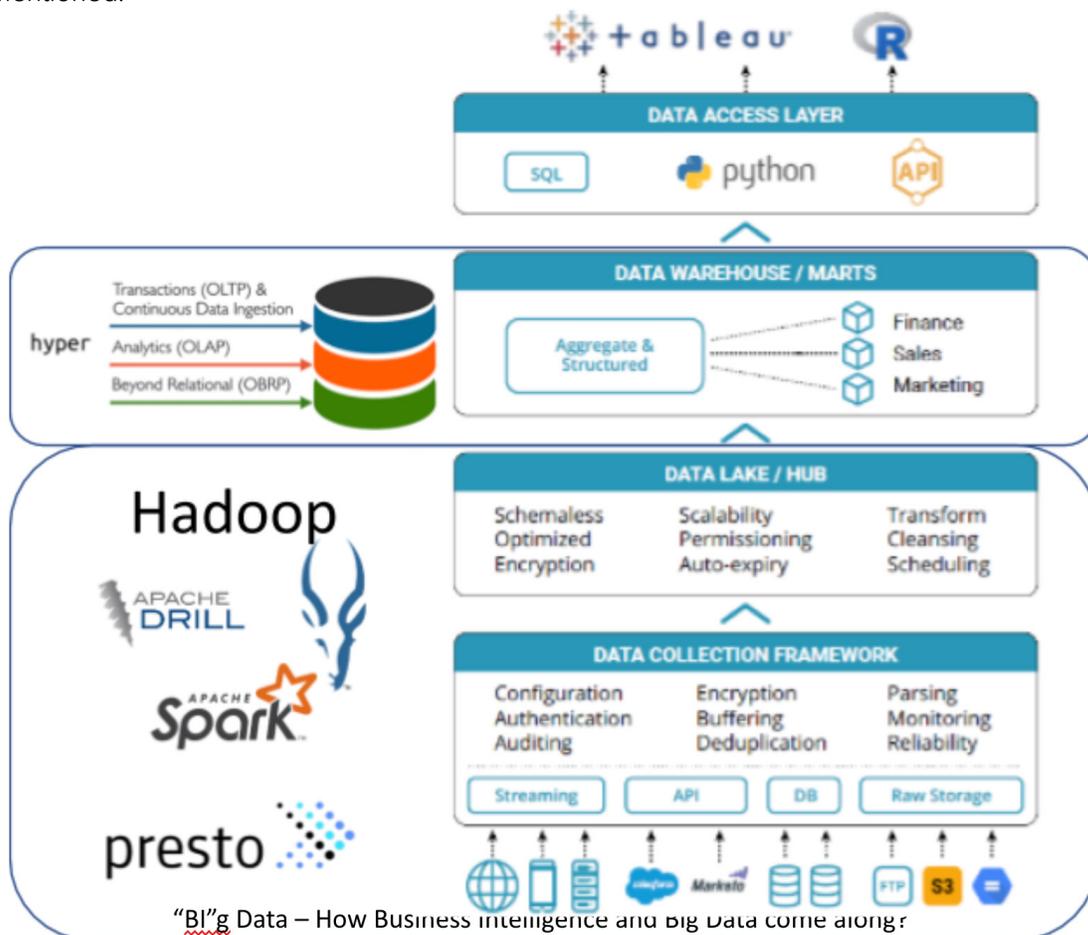


**sas** SAS uses analytic techniques, such as deep data profiling, to smartly identify data types that helps users target data transformation activities at relevant data.

SAS Research and Development is working to exploit the extensive analytics library at SAS to automatically identify potential sensitive information, to find data similarities, to propose data joining conditions, and to automate data quality transformations based on data analysis and user actions on the data.

**+ a b l e a u** Tableau integrates with Python and R for applying advanced analytic algorithms against data passed to Python and R from Tableau connected data sources like Hadoop.

Additionally, Tableau also has some built-in advanced analytics for clustering, forecasting, and regression that can be ran against data stored on Hadoop using any of the methods for connecting to Hadoop data previously discussed/mentioned.



## Authors



### Jeff Bailey

Jeff Bailey is the Data Access Product Management Technical Lead for data access and Hadoop. He is responsible for the SAS/ACCESS, In-Database, and Hadoop product lines. Jeff has been with SAS since 1992 and has spent most of his time helping customers use SAS with database management systems. He has written many papers covering how to use SAS with databases. In addition, he is frequently asked to speak on using SAS with databases. He has held positions in SAS R&D, Consulting, Education and Product Management. Jeff has a B.S. in Computer Science from the University of North Carolina at Charlotte.



### Cupid Chan

Cupid Chan is a seasoned professional who is well-established in the industry. His journey started out as one of the key players in building a world-class BI platform. He has been a consultant for years providing solutions to various Fortune 500 companies as well as the Public Sector. He is the Lead Architect for a contract in a government agency leading initiative on top of ~1PB data on both Big Data and traditional DB platforms. He is one of the 7-person Technical Steering Committee (TSC) and the Champion of the BI & AI SIG in ODPi.



### David Freriks

David Freriks is a Technology Evangelist in the Office of Strategy Management team at Qlik. Dave's mission is to help spread the word on the power of the Qlik Platform with Big Data systems and integration with partner technologies related to that ecosystem. He has spent 20+ years in the business intelligence space working at Qlik, SAP, IBM, and Cognos helping launch new products to market. Dave has a background in data warehousing and a Mechanical Engineering degree from Texas Tech University. He is married with two kids, and two Australian Shepherds.



### Benjamin Reyes

Benjamin Reyes is a Product Management lead at MicroStrategy Inc. in Tysons Corner, VA where he is responsible for the Enterprise Assets and Data Platform product portfolios. Ben has dedicated most of his professional career to the Enterprise Analytics space, contributing to the design and development of a broad range of platform capabilities such as data visualization, interactive data discovery and mobile BI amongst others.



### **Jason Tavoularis**

Jason Tavoularis is an Offering Manager for IBM Business Analytics. Over the last decade he has been engaging with IBM clients through roles in customer support, demonstrations and enablement, and product management. He has a bachelor's degree in Computer Engineering and an MBA from the University of Ottawa.



### **Gerard Valerio**

Coming up on 6-years at Tableau, Gerard Valerio currently leads a team of sales consultants and solutions architect in evangelizing Tableau to the U.S. Federal Government and helping customers see and understand their data using Tableau. Mr. Valerio is also an adjunct professor teaching Tableau at Montgomery College in Maryland and is the chief organizer of the Tableau Meet-Up for DC, Maryland, and Virginia. He has built a 20+ year career on data spanning mainframe and mid-range-based reporting systems (referred to as decision support systems and executive information systems) to first generation sub-Terabyte data warehouses in Oracle, Informix, and Sybase front-ended by business intelligence tools like SAP Business Objects, IBM Cognos, and MicroStrategy. Mr. Valerio also worked in the data integration space as a customer and employee of Informatica. His Big Data experience spans working Terabyte and Petabyte-sized data volumes staged within in-memory columnar databases like Vertica, Teradata, and others to structured/unstructured data residing in Hadoop-based data lakes to log data captured in Splunk. Mr. Valerio holds an Electrical Engineering degree from University of Illinois.