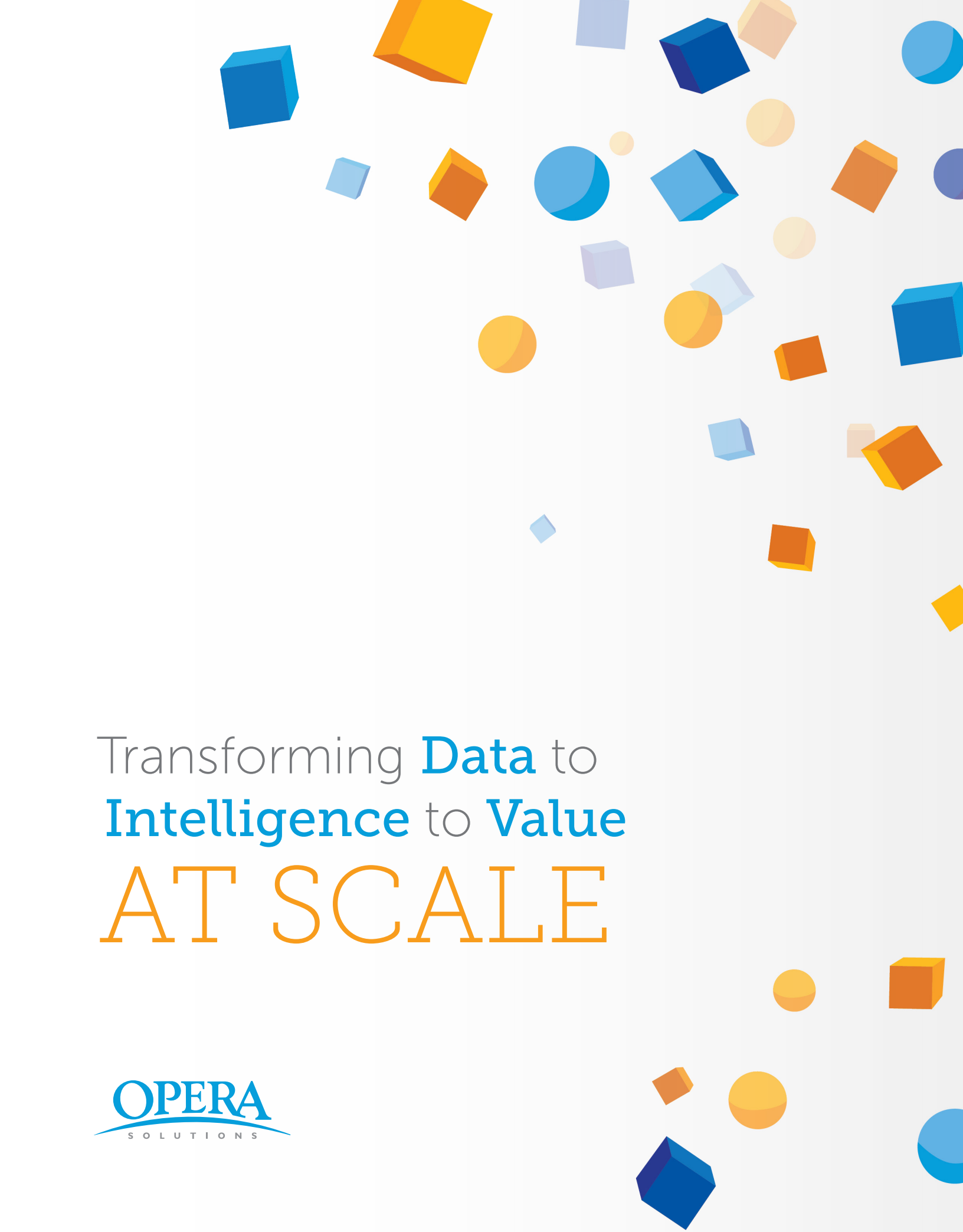


Transforming **Data** to **Intelligence** to **Value** AT SCALE

OPERA SOLUTIONS

OPERA SOLUTIONS

Transforming Data to Intelligence to Value At Scale

1 Introduction

As humans, we are constantly exploring our world in an attempt to extract meaning and value from our interactions with it. We have evolved to be able to utilize our five senses in an exquisite way; the number of raw data points being processed by our brains makes Big Data look tiny by comparison. We have mastered the process of extracting information from what would otherwise be meaningless impulses.

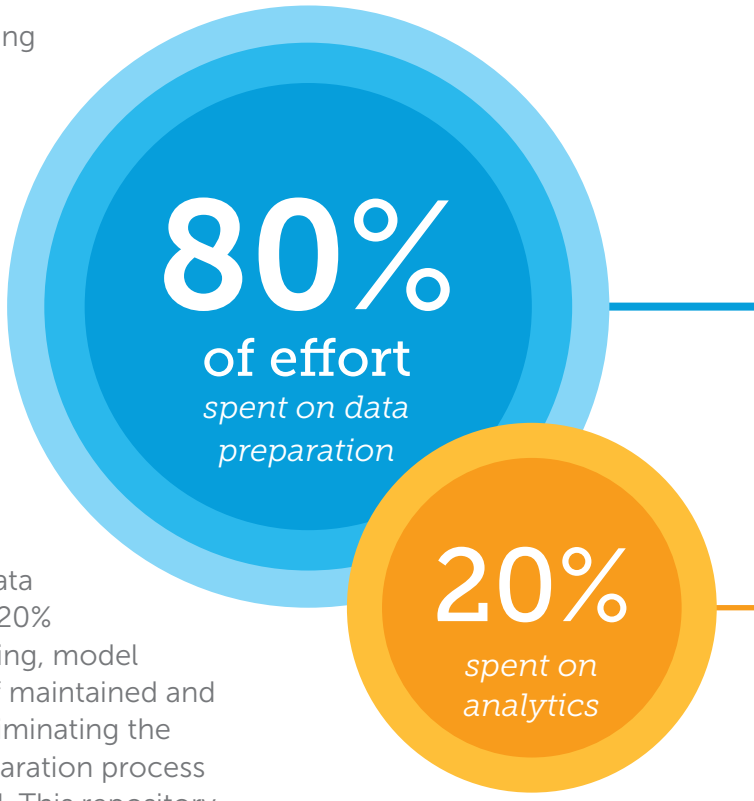
Big Data is a global phenomenon. Structured, semi-structured, and unstructured data points are being generated and captured at an ever-increasing pace.

But with state-of-the-art technology and data analytics, companies can extract value and insight from a myriad of raw data points. This white paper explores how enterprises are doing that today and why having and end-to-end software analytics platform is essential for companies to succeed in today's Big Data world.

2 The Sensible Journey From Data to Intelligence to Value

The current analytic approach to tackling Big Data starts with raw data and ends with intelligence, which is then used to solve a particular business use case so that data is ultimately translated into value. However, there is a crucial problem with the traditional approach being used to create analytic solutions: More often than not, the intelligence gathered from the data is not shared across the enterprise and is specific to solving a particular use case or business scenario. This approach is inefficient and wasteful at its core for the following reasons:

- Data and IT environments are too complex, making **use case implementation take too long.**
- Many use cases cannot be addressed because of strain on time and resources, and subsequently, **business opportunities are missed.**
- Hundreds of FTEs (full-time equivalents) are working on analytics, and **it is unclear if they are being utilized in the most productive way.**
- **There is an ever-increasing demand for analytics,** and the only current solution to scale is to add more FTEs.



In such an approach, 80% of the effort is spent on data preparation (cleaning, linking, processing), and only 20% is spent on analytics (BI, visualization, machine learning, model building). But now there's a way to provide a layer of maintained and refreshed intelligence (Signals) on top of the data, eliminating the need to go back to the raw data and repeat the preparation process every time a new use case needs to be implemented. This repository of Signals, along with the built-in science and technology that power it, is called a *Signal Hub*, and it is the single-most powerful platform in Big Data analytics today.

Signal Hub is highly scalable: There is scalability in processing large amounts of data as well as supporting the implementation of a myriad of use cases.

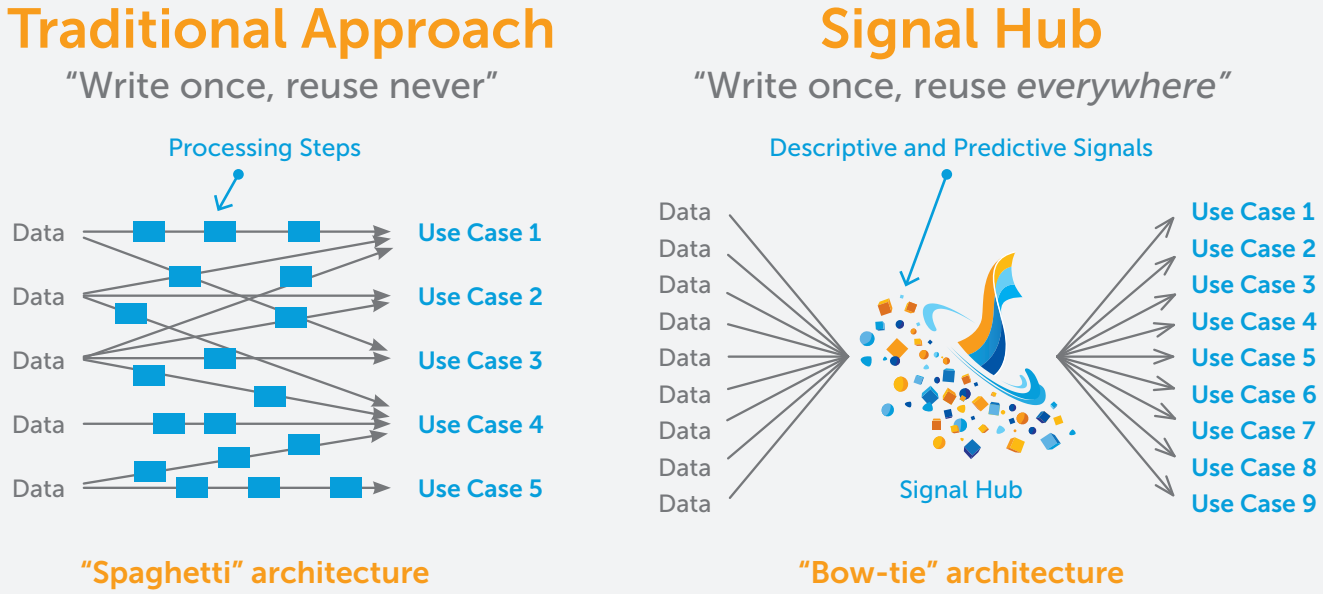


Figure 1: *Implementing analytics use cases in the traditional approach (left) and with Signal Hub (right).*

As you can see in the Traditional Approach in Figure 1, whenever a new use case is presented, users need to redo all the processing steps, starting with the raw data. Insights are trapped in silos, and there is simply no reuse of intelligence across different use cases.

With the Signal Hub approach, however, there's an intelligent layer between the raw data and use cases. This is where Signals are created and managed. Whenever a new use case needs to be implemented, data scientists can create them from existing Signals. A key observation here is that in a given domain, a lot of Signals are similar and reusable across different use cases. As the number of Signals grows, the model development time shrinks. In this "bow-tie" architecture, model developers are able to concentrate on creating the best predictive models because instead of starting every time with the raw data, they are able to use pre-developed Signals, ultimately decreasing time to value significantly for each use case.

As it would be expected, whenever a Signal is created in Signal Hub, it not only exists as a mathematical transformation but also as part of a semantic layer. This layer contains the metadata or taxonomy used to catalog all Signals inside Signal Hub. Semantic information allows users to easily browse through existing Signals. It lets them understand what each Signal represents, how they were created, and how they are being used. The metadata associated with a particular Signal gives it a context and ultimately promotes its reuse.

In short, Signal Hub is a revolutionary platform that enables organizations to operationalize the process of transforming data to intelligence, and then it maintains the intelligence as Signals in a production environment that allows the entire organization to access and exploit them for value creation.

2.1. What Are Signals?

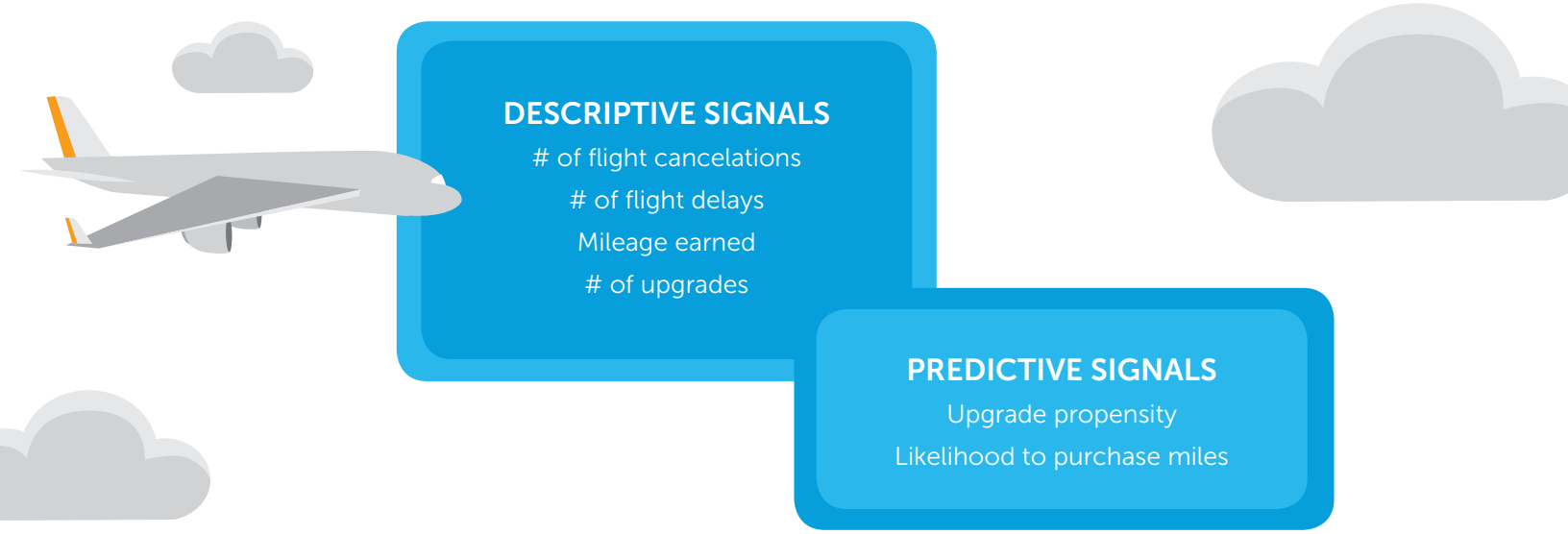
Signals are transformations applied to data. A transformation brings an aspect of the data into focus even if it was previously hidden. Signals can tell what happened in the past as well as what will happen next. While the past is captured by *descriptive Signals,* the future is represented by *predictive Signals.* Signal Hub enables the process of Signal creation.

Signals are useful information about events, customers, systems, and interactions. *They describe behaviors, events, and attributes of entities as well as predict future outcomes.*

The best way to explain Signals is by example. Since air travel is an experience most people have had, we will draw examples from the airline industry to illustrate the different levels of Signals enabled by Signal Hub.

If an airline is trying to improve customer satisfaction, it may want to know a great deal about the experience its customers are having when flying. For example, it may be important to find out if a specific customer had her last flight canceled. This Signal relies on flight information as it relates to customers. It becomes more powerful if it looks at the same data over a period of time. For instance, it could look at the total number of flight cancelations a given customer experienced over the previous 12 months.

Such a Signal could help measure levels of satisfaction. Others that could feed into a satisfaction score would be how many times a customer was delayed or upgraded in the past one year. Descriptive Signals such as these can also look across different data domains to find information that may make the case for a customer to continue flying with the same airline. For example, a Signal may identify a partner hotel a customer tends to stay with so that a combined discounted deal, including airline and the same hotel brand, can be offered to her. This allows for airlines to benefit from her contentment level with the specific hotel partner. Note that in this case, raw input data is consolidated across industries to create a specific relationship with this particular customer.



Descriptive Signals also benefit from Signal Hub's ability to easily concatenate events over time. And so a flight cancelation followed by a hotel stay indicates that the customer got to the destination but with a different airline or some other mode of transportation.

While descriptive Signals are a key driver for many decisions, Signal Hub takes analytics to a whole new level by enabling deep analytics via the creation of predictive Signals. Predictive Signals tell the likelihood of a given event happening in the future. These are the crème de la crème of Signal Hub's repertoire of capabilities. A predictive Signal is usually created with a use case in mind. It can be as simple as computing the affinity associated with a particular customer as related to a segment of people who tend to fly on red-eye flights or as complex as computing the propensity level for the same customer to upgrade to business class. Predictive Signals allow for the enterprise to determine what a customer will do next or how she'll respond to a given communication and then plan appropriately.

Inside Signal Hub, Signals can also be created by combining predictive and descriptive Signals. For instance, it can identify high-yield customers who have a high propensity (predictive) to buy a discounted ticket to destinations that are increasing in popularity (descriptive).

2.2. Enabling the Journey to Higher Productivity

As the enterprise moves forward with Signal creation inside Signal Hub, it turns the traditional analytics approach on its head. The more use cases that are executed using Signal Hub, the less time it takes to implement them over time. That's because the answers to a problem may already exist inside Signal Hub after a few rounds of Signal creation and use case implementation.

This new Signal-based approach enables teams to "write once and reuse everywhere," as opposed to the traditional approach, which falls into the trap "write once and reuse never."

The Signal Hub approach transforms the industry landscape as follows:

Traditional Approach	**Signal Hub** Approach
Write once and reuse ***never.***	Write once and reuse ***everywhere.***
Analytics are done ***only by data scientists.***	Analytics are done by ***data scientists and business users.***
Insights are ***trapped in silos.***	Insights are ***shared across the enterprise and made available for reuse*** for any use case.
Scientists spend ***80% of their time performing data preparation steps*** and ***20% on analytics.***	After the initial use case is complete, scientists ***skip the 80% of work*** that would otherwise be repeated for each subsequent use case and focus on analytics instead. ***Signal Hub flips the data preparation/analytics ratio from 80/20 to 20/80.***

Making this possible are the three major components on which Signal Hub is built. As shown in Figure 2, these are the Knowledge Center (KC), Integrated Development Environment (IDE), and Signal Hub Server. All of these work together to realize the Signal-based approach. KC facilitates the transformation of intelligence to value through the exploration and consumption of Signals. The IDE enables scientists to more effectively transform data to intelligence through the creation of Signals. And the Signal Hub Server executes analytics by running the code and producing the Signal output.

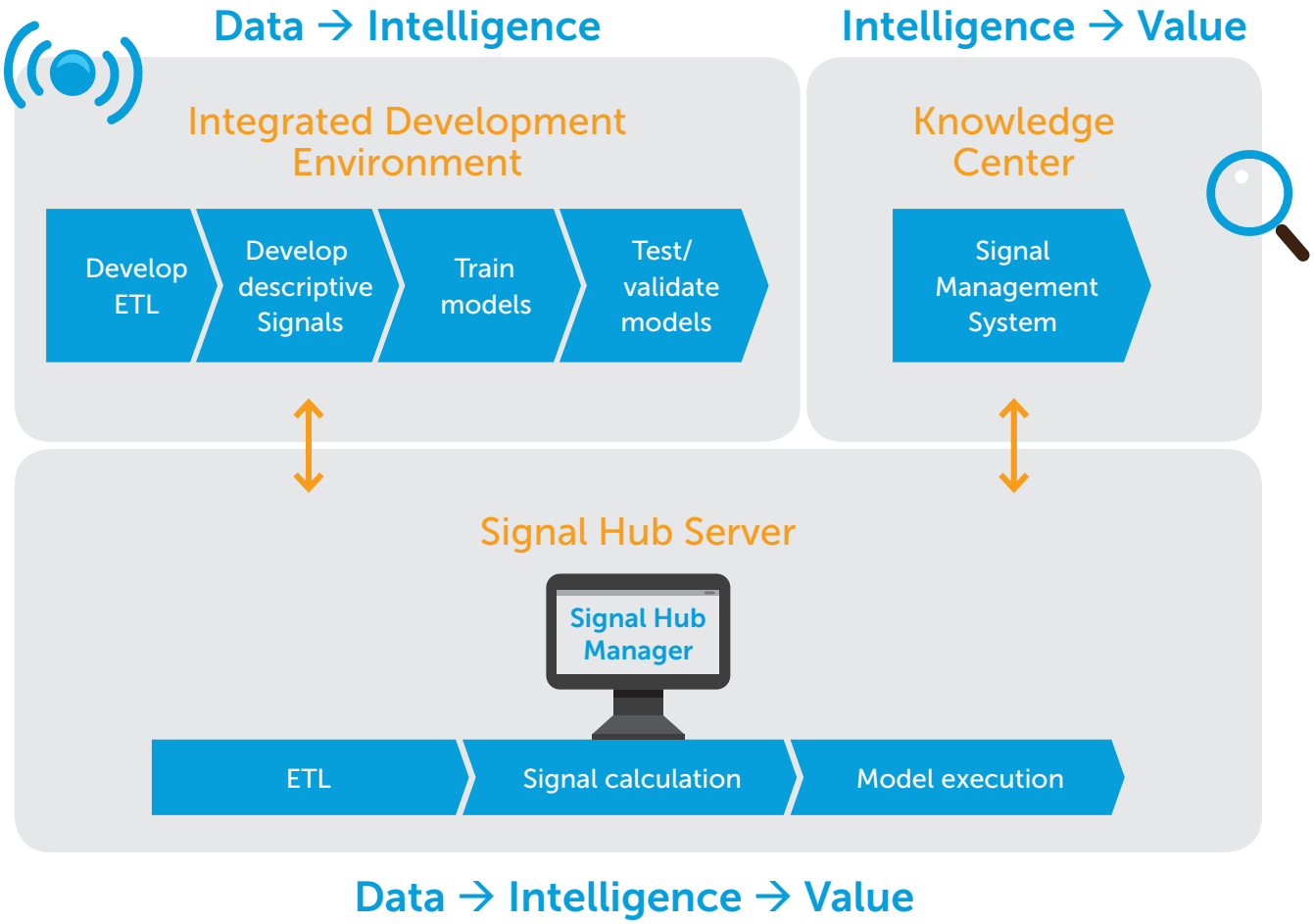


Figure 2: *Signal Hub components: IDE, KC, and Signal Hub Server.*

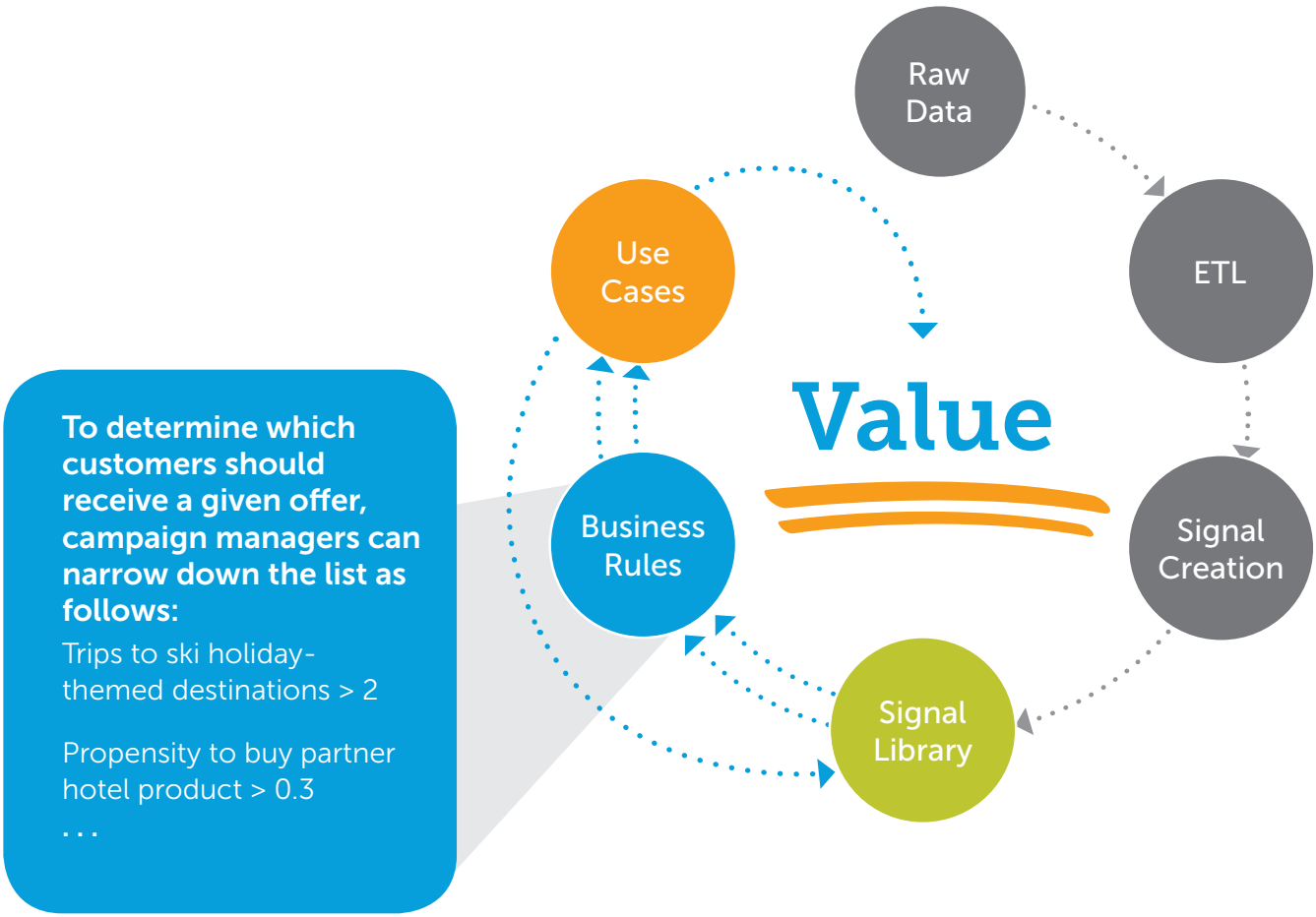
2.2.1. Accessing Intelligence via the Knowledge Center

As an integral part of Signal Hub, KC was designed from the ground up as an interactive Signal management system. It enables model developers and business users to easily find, understand, and reuse Signals that already exist in the Signal Library inside Signal Hub. KC allows for the intelligence (Signals) to be accessed and explored across use cases and teams throughout the enterprise. Whenever a new use case needs to be implemented, KC enables relevant Signals to be reused so that their intrinsic value naturally flows toward the making of a new analytic solution that drives business value.

Multiple features of KC facilitate this revolutionary way of accessing and consuming intelligence. The first is its filtering and searching capabilities. When Signals are created, they are tagged based on metadata and organized around a taxonomy. KC empowers business users to explore the Signals through multiple filtering and searching mechanisms. Below are the key components of the metadata in each Signal:

1. Business description, which explains what the Signal is (such as the number of times a customer sat in the middle seat on a long-haul flight in the past three years).
2. The Opera Solutions–developed taxonomy, which shows each Signal's classification based on its subject, object, relationship, time window, and business attributes. For example: subject = customer, object = flight, relationship = count, time window = single period, and business attributes = long haul and middle seat.

KC users are then able to explore and identify Signals based on this metadata when executing use cases by using filtering and free-text searching. So, for example, say a marketing campaign manager for an airline sees that for the upcoming month, partner ski resort hotels are 70% vacant, and planes to ski-themed destinations are 50% vacant. She wants to fill those vacancies by offering a free sixth night for a five-night trip along with a 20% discount on airfare. She could use KC to search for customers that have a strong preference for skiing, are price-sensitive, are comfortable with a short lead time to booking, and have a high likelihood of purchasing a hotel product.



With KC, she could readily identify relevant existing Signals, apply filters to those Signals, and export a list of customers. Compare that with the legacy approach, which would take four to six weeks.

KC also allows for a complete visualization of all the elements involved in the analytical solution. Users can visualize how data sources connect to models through a variety of descriptive Signals, which are grouped into Signal Sets depending on a prespecified and domain-driven taxonomy. The same interface also allows users to drill into specific Signals.

Once a Signal of interest is identified, users can gain a deeper understanding of the Signal by exploring its lineage from the raw data through all transformations, providing insight into how a particular Signal was created and what the value truly represents.

Additionally, users can isolate exactly which columns in the raw data or other Signals were combined to create the Signal of interest. Finally, they can identify which Signals, if any, consume the Signal of interest and view the code that was used to define it. All of this capability serves multiple purposes:

- Provides a better understanding of Signals
- Helps scientists determine what codes need to be evoked in the production system to calculate the Signal
- Makes Signal management easier and faster

KC contains visualization capabilities to allow users to explore the values of Signals directly in the Signal Hub platform. Users are also able to apply business rules to Signals to filter the data and target subsections of the population, as shown in Figure 3, Rule Sets 1 and 2. This is particularly important, as it enables business users to build sophisticated prescriptive models, allowing for true democratization of Big Data analytics across the enterprise.

2.2.2. Integrated Development Environment — Signal Hub's IDE

The IDE is the workbench of Signal Hub. Together with the KC, it enables users to interact with all the functionality and capabilities offered in Signal Hub via a rich graphical user interface (UI).

The IDE is intrinsically an environment to develop end-to-end analytic solutions. It allows all the components of the entire analytic modeling process to come together, from data to Signals. It also provides an environment for the coding and development of the following:

1. Data schemas
2. Visualization and maintenance of staging, input, and output data models
3. Data quality management processes (e.g. missing value imputation and outlier detection)
4. Collections (the gathering of raw data files with the same data schema)
5. Views (logic to create a new relational dataset from other views or collections)
6. Descriptive and predictive Signals
7. Model validation and visualization (measuring of model performance through ROC, KS, Lorenz curves, and other methods)

Both data models and schemas can be developed within the IDE or imported from popular third-party data modeling tools such as CA Erwin. The data models and schemas are stored along with the code and can be governed and maintained using modern software lifecycle tools. Typically, at the beginning of a Signal Hub project, the IDE is used by data scientists for profiling and schema discovery of unfamiliar data sources. Signal Hub provides tools that can discover schema (i.e., data types and column names) from a flat file or a database table. It also has built-in profiling tools, which automatically compute various statistics on each column of the data such as missing values, distribution parameters, and frequent items. These built-in tools accelerate the initial data load and quality checks.

Once data is loaded and discovered, it needs to be transformed from its raw form into a standard representation that will be used to feed the Signals in the Signal Layer. Using the IDE, data scientists can build workflows composed of "views" that transform the data and apply data quality checks and statistical measures. The Signal Hub platform can continuously execute these views as new data appears, thus keeping the Signals up to date.

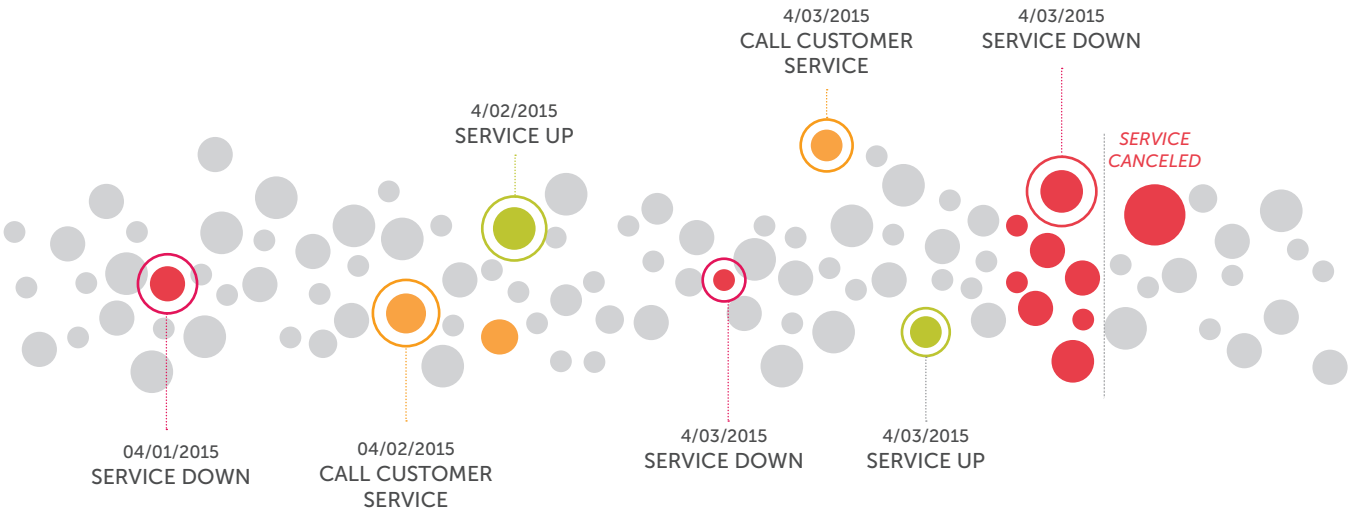
2.2.2.1 Analytics Through Signal Creation and Computation

The IDE possesses amazing capabilities for ingesting and manipulating data, but its main proposition is on enabling data scientists to extract intelligence from data through Signals. It does that by offering scientists a workbench in which the past (historical data) is used to create descriptive Signals, which are then used to create predictive Signals. In this way, the IDE can be thought of as a Signal-creation machine.

Descriptive Signals and the Signal API

Opera Solutions created the Signal API (application program interface) to allow data scientists to veer away from the implementation details and focus solely on data analysis, thus maximizing productivity and code reuse. The Signal API editor provides several features that help reduce errors and speed up use case development and Signal creation. For example, in the API, scientists can access an ever-growing set of mathematical transformations that allow for the creation of powerful descriptive Signals, along with a syntax that is clear, concise, and expressive.

It also allows for easy implementation of complex pattern-matching Signals. Take the example below, for instance: In the telecom industry, one pattern could be a sequence of events that are relevant for measuring attrition, such as a service disruption followed by one or more customer complaints, followed by restored service. One would need to use complex code to come up with the same functionality for "connecting the dots" in other more traditional ways.



The Signal API also provides a direct link between the IDE and KC. With it, users can add metatags and descriptions to Signals directly. These tags and taxonomy information are then used by KC to enable Signal search and reuse, enabling a "write once, reuse everywhere" mentality, which greatly enhances productivity.

Predictive Signals

As for predictive Signals, training and testing of models can easily be done in the IDE through its intuitive and interactive UI. Current techniques available for modeling and dimensionality reduction include SVMs, k-means, decision trees, association rules, linear and logistic regression, neural networks, RBM, and PCA. This list also includes Deep AutoEncoder, which allows data scientists to train and score deep learning nets. Opera Solutions' scientists have amassed great expertise in using advanced machine-learning techniques, such as Deep AutoEncoder and RBM, to project data from a high-dimensional space into a lower-dimensional one. These techniques are then used together with clustering algorithms to understand customer behavior.

Interoperability and PMML

Through the IDE, it is also extremely convenient to export descriptive Signals into a flat file for the training of predictive models outside Signal Hub. When the model is ready, it can then be brought back to Signal Hub via the PMML (Predictive Model Markup Language) standard. This feature is very useful if a specific machine-learning technique is not yet part of the model repertoire available in Signal Hub. It also allows Signal Hub to ingest models created by our clients in third-party analytic tools (including R, SAS, and IBM SPSS). The use of PMML allows Signal Hub users to benefit from a high level of interoperability among systems where models built in any PMML-compliant analytics environment can be easily consumed.

2.2.3. Signal Hub Server

The Signal Hub Server is responsible for the end-to-end processing of data and its refinement into Signals. It is also responsible for making Signal Hub a truly horizontal platform, as it enables users to solve problems across industries and domains.

Not all business problems require Big Data solutions, but the volume of data being produced and captured is constantly expanding. The Signal Hub Server is able to perform large-scale processing of terabytes of data across thousands of Signals. It follows a data flow architecture for processing on a Hadoop cluster. Hadoop 2.0 introduced YARN (a large-scale, distributed operating system for Big Data applications), which allows many different data processing frameworks to coexist and establishes a strong ecosystem for innovating technologies.

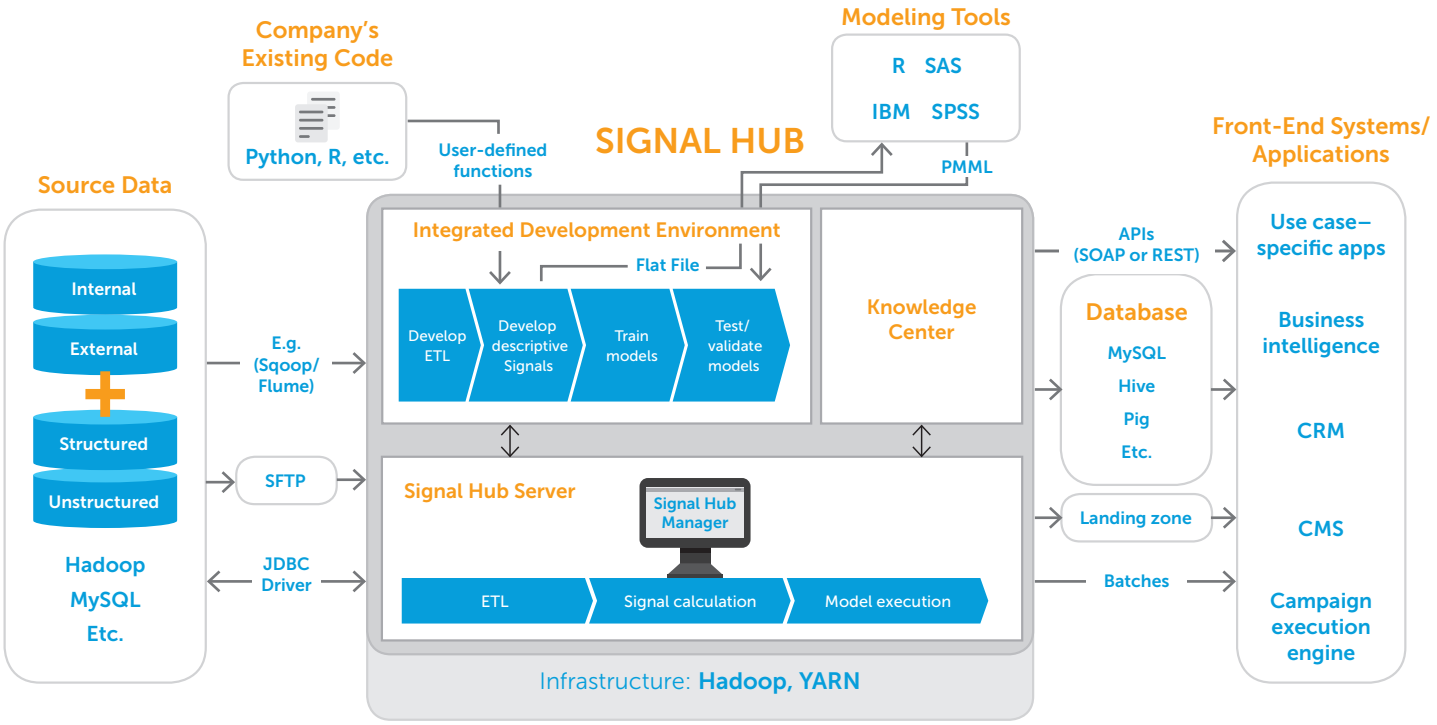


Figure 4: *Signal Hub allows for easy and fast extraction of intelligence from data. The Signal Hub Server enables such a process by supporting the following:*

1. *Data ingestion of various formats and sources*
2. *The execution of all the analytical steps implemented in the IDE*
3. *The Signal management capabilities of KC*
4. *The consumption of Signals by front-end systems*

Signal Hub is YARN-compatible, and therefore all Signal Hub solutions are automatically Hadoop-certified and can be managed and administered alongside other applications. Enterprises that use Signal Hub can leverage their investment in Hadoop technologies and IT skills and run Signal Hub alongside their current Hadoop applications.

The platform architecture provides great deployment flexibility. It can be implemented on a single server as a single process — using so few resources that code development can take place on a laptop — or it can run on a large-scale Hadoop cluster with distributed processing, without modifying any code. This allows scientists to develop code on their own computers and then move it into a Hadoop cluster to process large volumes of data. The Signal Hub Server architecture addresses the industry need for large-scale production-ready analytics — a need that popular tools such as SAS and R cannot fulfill even today, as their basic architecture is fundamentally main memory-limited.

The Signal Hub Server

integrates seamlessly with a variety of front-end systems and ensures that existing investments in analytics can be reused with no need for recoding.

As shown in Figure 4, the Signal Hub Server allows companies to absorb information from various data sources to be able to address many types of problems. It can ingest both internal and external data as well as structured and unstructured data. As part of the Hadoop ecosystem, Signal Hub Server can be used together with tools such as Sqoop or Flume to digest data after it arrives in the Hadoop system. Alternatively, the Signal Hub Server can directly access any JDBC (Java Database Connectivity)-compliant database or import various data formats transferred (via FTP) from source systems.

The Signal Hub Server integrates seamlessly with a variety of front-end systems. Signals are easily fed into BI tools (e.g. Pentaho, Tableau), CRM systems, and campaign execution engines (e.g. HubSpot, Exactarget). Data is transferred in batches, written to a special data landing zone, or accessed on-demand via APIs. The Signal Hub Server also integrates with existing analytic tools, pre-existing code, and models. Client code (e.g. Python, R, C++) can be loaded as an external library and executed within the server. As mentioned before, a model developed in SAS, R, or SPSS can be consumed and run within Signal Hub via the PMML standard. All of this ensures that existing investments in analytics can be reused with no need for recoding.

The Signal Hub Server automates the processing of inputs to outputs. Because of its data flow architecture, it has a speed advantage; benchmarking the server data processing performance showed a 10–15X improvement over MapReduce. The Signal Hub Server has multiple capabilities to automate server management. It can detect data changes within raw file collections and then trigger a chain of processing jobs to update existing Signals with the relevant data changes.

The Signal Hub Server also provides a management and monitoring console, dubbed the Signal Hub Manager, for analytics operational stewards (from IT, business, and science). With it, they can understand and manage the production quality and computing resources of the deployed Signal Hub.

3 | Summary

In Signal Hub, internally available enterprise data as well as external data come seamlessly together to allow for a holistic view of customers and processes through the implementation of Signals. Its three components — Knowledge Center, Integrated Development Environment, and Signal Hub Server — work together to enable this functionality.

The KC is a centralized place for institutional intelligence and memory. It enables the management and reuse of Signals, which leads to scale and an explosion of productivity. With KC, business and analytic users alike can view analytics in a revolutionary way, opening up what has historically been a black box. Intelligence that was once stored on individuals' laptops, in their heads, or lost within code is captured in Signals and organized and exposed to everyone for consumption.

The IDE, for its part, is an integrated productivity tool for data scientists and developers, offering analytic functionalities and approaches for the making of a complete analytic solution, from data to intelligence to value.

Finally, the Signal Hub Server ties it all together under the hood by providing fast and scalable processing of data, code, and artifacts in Hadoop via a data-flow execution engine.

Signal Hub industrializes the analytics process and flips the data preparation/analytics ratio from 80/20 to 20/80. It integrates data from a variety of sources, which enables the process of Signal creation and utilization by business users and systems. This leads to the democratization of insights across the enterprise. A cascade of events is then set in motion in which data truly creates a positive impact throughout the entire business ecosystem, making it consistent and optimized. In this way, Signal Hub creates a solid path for a company's journey from data to intelligence to value.

For more information, please visit our website, **www.operasolutions.com,** email **interest@operasolutions.com,** or call **1-855-OPERA-22.**

OPERA SOLUTIONS

Jersey City | Boston | San Diego | London | Shanghai | New Delhi

ABOUT OPERA SOLUTIONS, LLC

Opera Solutions is a global provider of advanced analytics software solutions that help organizations extract actionable insights from Big Data faster, more efficiently, and with greater business impact than traditional approaches. Signal Hub™, the company's flagship technology platform, uses data science to transform data to intelligence and intelligence to value. Signal Hub and its family of associated enterprise solutions generate insights that power critical strategic and operational activities across multiple functional domains for enterprises and governments worldwide. The company has offices in North America, Europe, and Asia. For more information, visit www.operasolutions.com.