As the growth in deployment of massive sensor networks by enterprise and government organizations continues to redefine the sensor-to-insight data flow, it is easy to believe that existing Big Data software can be used to build applications involving fast, streaming data. However, realizing value from larger amounts of fast data requires a different solution than those for deriving value from data-at-rest. This is because data stored using Hadoop-based storage are not fast enough to act on real-time data.

Apache Hadoop is a scalable, open-source data storage platform that has emerged as the backbone of a growing number of enterprise organizations’ Big Data systems. Hadoop consists of two major components: the Hadoop Distributed File System (HDFS) for storing data and a processing framework to enable parallel computation on the same computing nodes to minimize data movements.

In the initial version of Hadoop, MapReduce was the primary computing framework, but it has since been extended with YARN and Spark to support a variety of processing models and leverage in-memory capabilities, respectively. Two of the key contributions of the Hadoop ecosystem were the support of a schema-on-read approach to facilitate the ingestion of a variety of data types and a software framework for leveraging large numbers of commodity servers for scale-out computing.

Sensor networks generate massive amounts of fast, streaming data. Some have proposed piecing together Hadoop’s capabilities to ingest and persist massive quantities of data together with high throughput messaging tools, such as Kafka, to address the challenges of building Internet of Things (IoT) applications. Apache Kafka is an open-source, fast, scalable, durable, and fault-tolerant publish-subscribe messaging system. Kafka is often used in situations requiring high throughput, reliability and replication.
Challenge of Improving Sensor-to-Insight Data Flow

The idea of composing solutions from software components, such as Hadoop and Kafka, to support IoT applications is appealing on the surface, but the reality involves greater complexity. Tools like Kafka are designed to only move Fast Data into HDFS, and they do not perform any processing during the data movement.

This approach results in "dumping" great volumes of raw sensor data in HDFS for analysis after the fact. This limits the system’s ability to react to real-time events, as well as requires far larger computational resources to analyze the massive volumes of data after they are loaded into HDFS.

A more powerful ingestion system involving real-time data transformation and integration leads to better data for HDFS, because data can be filtered, aligned, aggregated, and enriched during the ingestion process. Data can also be queried directly during the ingestion rather than "after the fact" to detect anomalies and other patterns of interest.

A more streamlined Fast Data pipeline also enables the combination of transactions with analytics to facilitate faster development of complex IoT applications. For example, the nature of many sensor data can benefit from using machine-learning methods to automatically transform Fast Data. An advanced data ingestion system can “fuse” streaming data from sensors with static data from analytic tools based on machine-learning to automatically correct and transform sensor data for additional analysis.
**Big Data in Real Time**

By leveraging Hadoop with Apache Spark and supporting streaming messaging tools, such as Kafka, atop an object-oriented database that is purpose-built for information fusion, Objectivity's ThingSpan can help companies achieve business insights from Big Data and real-time streaming data with a high degree of efficiency at scale.

Information fusion is the process with rich pedigree for dealing with association, correlation, and combination of data from multiple sources to achieve refined estimates of parameters, events, and behaviors of observed entities.

ThingSpan ensures superior performance by organizing data about people, locations, events, and devices into a logical model involving objects and the relationships between them. This allows for enriched and transformed data, as well as associated metadata to be persisted, to simplify the support of complex, multi-dimensional queries associated with IoT applications and analytics.

ThingSpan's object-oriented approach to information fusion makes it faster and easier to create systems capable of managing data volumes well beyond the petabyte level. Now organizations can transform Big and Fast Data from generic to relevant in real time, thereby maximizing business value.

**About Objectivity, Inc.**

Objectivity, Inc. is a pioneer in high-performance distributed database platforms that power mission-critical applications for the most demanding and complex data sources in the enterprise. Objectivity enables organizations to accelerate time-to-value of their data assets at scale by enriching Big Data with Fast Data.

With a rich history serving Global 1000 customers and partners, Objectivity holds deep domain expertise in fusing vital information from massive data volumes to capture new revenue opportunities, drive competitive advantages, and deliver better business value. Objectivity is privately held with headquarters in San Jose, California. Visit www.objectivity.com to learn more.